

Article

Enhancing Causal Text Detection Using Uncertainty-Weighted Machine Learning Ensembles

Sivachandra K B ¹, Neethu Mohan ¹, Mithun Kumar Kar ¹, Sikha O K ² and Sachin Kumar S ^{1,*}¹ Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, Amrita University, Ettimadai 641112, India; cb.sc.p2dsc24018@cb.students.amrita.edu (S.K.B.)² BCN MEdTech, Department of Information and Communication Technologies, University of Pompeu Fabra, 08002 Barcelona, Spain

* Correspondence: s_sachinkumar@cb.amrita.edu

Abstract

Causal inference in text data has been a demanding objective in the field of natural language processing, mainly due to the intrinsic ambiguity and context sensitivity inherent in data, inducing uncertainty. Diminishing this uncertainty is essential in identifying reliable causal connections and advancing predictive consistency. In this research, we introduce an uncertainty-aware ensemble architecture that combines multiple text embedding schemes with both linear and nonlinear classifiers to boost causal text detection. Both sparse and neural-level embeddings were employed, and then combined it with an ensemble weighting approach based on two uncertainty estimation techniques, namely entropy-based and KL divergence-based. Unlike conventional ensemble methods with uniform or fixed voting strategies, our approach assigns weights inversely proportional to classifier uncertainty, ensuring that confident models exert greater influence on the final decisions. Our results show that TF-IDF, through its effective word frequency weighting scheme, consistently outperforms other embedding techniques, achieving better performance across both linear and nonlinear classifiers on both datasets (News Corpus and CausalLM–Adjective group). The experimental results show that our uncertainty-aware ensemble approach enhances both calibration and confidence predictions. Entropy-based weighting improves confidence in the case of linear classifiers with accuracy, F1-score, entropy and prediction confidence values of 94.3%, 94.0%, 0.382 and 0.774, respectively, while in the case of nonlinear classifiers the KL divergence-based weighting acquires a better performance with an accuracy of 97.6%, F1-score of 97.2%, KL Mean value of around 0.055 and LogLoss of 0.221.

Keywords: causality; word embedding; machine learning; uncertainty; ensemble learning

Academic Editors: Rui Zhang and Olga Kurasova

Received: 7 October 2025

Revised: 2 February 2026

Accepted: 3 February 2026

Published: 2 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Causality is a vital area in human decision-making and reasoning, enabling us to differentiate true cause-and-effect relationships from mere correlations. In recent years, research on causality detection from textual data has garnered significant focus, particularly with the expansion of digital information resources, including news articles, scientific publications, medical records, and social media streams. Interpreting cause–effect relations within text corpora is crucial in a diverse range of applications, mainly in discovering biomedical findings, policy analysis, and risk evaluation in finance and clinical settings. Unlike surface-level NLP tasks, causal inference provides a deeper interpretive understanding, empowering systems not only to recognize events but also to reveal the underlying

process that connects them. Despite its significance, causality recognition in textual data still constitutes a difficult issue. Linguistic expressions often convey causal meanings in indirect, direct, unclear, and context-sensitive forms, leading to limitations for traditional machine learning models in learning causal relationships [1,2]. Traditional techniques, including rule-based or handcrafted-feature approaches, often have a shortcoming in scaling across diverse domains. Due to their model-agnostic characteristics, deep learning models, even when achieving high performance, are often less sensitive to domain-specific features. As a result, they commonly face issues such as data insufficiency, domain drift, and poor calibrated results [3,4]. Thus, a key component of causal analysis research lies in measuring prediction uncertainty and incorporating this information into decision-making.

Uncertainty quantification has become prominent as a foundation for developing reliable NLP systems [5]. In this research, we aim to emphasize classifier-level uncertainty—quantifying how confident or uncertain models are when identifying causal relations in text data, rather than modeling linguistic-based ambiguity. Most of the earlier research focuses on increasing the evaluation scores; however, the current work focuses on uncertainty-based weighting. Employing uncertainty not only increases the interpretability of prediction but also helps to reduce the risk of error propagation in downstream decision workflows. Beyond this, uncertainty quantification offers a principled approach to identify both underfitting and overfitting—increased model uncertainty often suggests limited model capability or poor generalization, while miscalibrated confidence predictions signal overfitting. These issues often arise in ML techniques for text, where simple linear models miss to extract complexity and thus can underfit complex causal patterns, and more powerful nonlinear learners may overfit by learning dataset-specific cues rather than true causal structure. In causal relation detection, where incorrect predictions can often lead to incorrect reasoning chains, uncertainty-aware modeling is essential for improved classification and trustworthiness. To reduce these challenges, an ensemble framework has shown effective by combining the capability of diverse models to achieve more stable and generalizable predictions. Traditional ensembles often assign uniform weights to all classifiers or rely on static weighting strategies, neglecting that individual models display varying confidence levels across samples. This motivates the adoption of uncertainty-weighted ensemble learning, where the classifiers' influence is dynamically adjusted based on their uncertainty. In this framework, models producing confident predictions have a strong influence on the final classification task, while a model with less confident predictions contributes less, ensuring a trade-off between accuracy and trustworthiness.

In this research, we put forward an uncertainty-guided ensemble framework for recognizing causal relations in text data by integrating a wide range of embedding techniques—TF-IDF, Word2Vec-CBOW, Word2Vec-SkipGram, GloVe, and FastText—paired with various classifiers, such as Support Vector Machines, XGBoost, Random Forest, and Extra Trees. Our evaluations are performed on two benchmark datasets: the Causal News Corpus (CNC) and the CausalLM-Adjective subset. The CNC consists of short, event-focused news sentences containing direct and indirect causality markers, while the CausalLM dataset consist lengthier, evenly distributed samples in which causality is often expressed through adjective-based causal cues. These datasets differ significantly in text length, vocabulary diversity, and the occurrence patterns of causal markers, and these attributes guide the choice of embedding techniques and in designing ensemble architecture. This heterogeneous setup integrates weak learners (e.g., linear and shallow models) with powerful learners (e.g., tree-based ensembles capturing complex nonlinear interactions), allowing mutually supportive learning behaviors that are particularly useful in limited-resource environments. FastText is used due to its character-level modeling via character n-grams, which provides stability for rare, complex, and misspelled tokens—a benefit that is absent in word-level

embeddings such as Word2Vec. To improve reliability, the system computes uncertainty via entropy along with KL divergence disagreement metrics and combines them to develop an uncertainty-aware weighted ensemble model, in which the classifier influence is inversely scaled by its uncertainty. This weighted ensemble framework provides improved consistency, transparency, and predictive performance for causal relation detection in comparison with standalone models and uniformly weighted ensembles.

In this research work, we mainly concentrate on:

- Highlighting the importance of uncertainty estimation in causal detection tasks in text and demonstrating its impact in improving reliability.
- Developing entropy- and KL divergence-based weighting ensembles, allowing confident classifiers to exert stronger influence on final predictions.
- Performing extensive experiments across diverse embeddings (TF-IDF, BoW, Word2Vec-CBOW, Word2Vec-Skip-Gram, GloVe, and FastText) and classifiers (linear and non-linear), evaluating performance on benchmark datasets including the Causal News Corpus and CausalLM.
- Showing that uncertainty-weighted ensembles provide consistent improvements in both accuracy and calibration, offering a generalizable and reliable approach for causal detection in textual data.

2. Literature Survey

Research on causal relation detection in natural language processing (NLP) has advanced rapidly, integrating knowledge from linguistics theory, ML modeling, and causal inference. Traditional approaches depend on rule-based patterns to identify clearly expressed causal text, offering transparency but lacking scalability and showing weakness in detecting implicit or context-dependent causal relations. The introduction of neural representation learning, methods based on distributed embeddings such as Word2Vec, GloVe, and FastText, have enabled the modeling of semantic relations through contextual co-occurrence rather than explicit meaning. A variety of classifiers—including support vector machines, tree-based ensembles, and neural architectures—have since been applied to this task. However, most prior work has emphasized improving accuracy while giving limited attention to model reliability under uncertainty. Recent developments in uncertainty estimation and ensemble learning show that incorporating predictive confidence can significantly improve robustness. Positioned within this context, our work unifies causal detection, uncertainty modeling, and ensemble learning into a single framework that extends the existing contributions.

2.1. Survey on Causality Text Detection

Causal detection research has been shaped significantly by the evolution of available open-domain (OD) datasets. Early work relied on small, domain-specific corpora annotated with explicit causal connectives (e.g., because and therefore), which constrained the generalizability of the resulting models [6,7]. As the field matured, researchers introduced larger and more diverse resources, including newswire datasets such as SemEval-2010 Task 8 [8], biomedical literature collections, and social-media-based corpora, enabling studies of both explicit and implicit causality. Because these datasets differ widely in annotation schemes, domains, and causal relation types, designing a single model that performs robustly across them remains challenging. To resolve these limitations associated with human annotation, several modern approaches have introduced hybrid annotation pipelines such as semi-automatic, distant-supervision, and weak-labeling strategies to improve the scalability of causal relation extraction [9,10]. In complementary direction, syntactic and dependency-based techniques—enabled by tools such as UDPipe [11,12], have shown that structural

cues can contribute to the identification of latent causal patterns. In our study, we evaluate the proposed uncertainty-weighted ensemble framework across these benchmark datasets to ensure that its performance reflects realistic and diverse causal detection scenarios. The challenge of extracting cause–effect relationships from news text has achieved significant growth with the publication of the *Causal News Corpus* (CNC) by Tan et al. [13]. Their study offers extensive annotations of causal relations within event-centric contexts, providing a standard for span-level evaluations and relation classification tasks. The authors underline key challenges, namely implicit causality, long-range dependencies, and unclear expression in causal cues, which have encouraged subsequent methodological advancements in representation learning and extraction techniques. The use of bidirectional encoder networks from the standard transformer was employed for the classification task. Patel [14] expanded this research direction by developing a *supervised contrastive learning* model for classifying causal events. His framework clusters semantically alike causal links within the embedding space while distancing unrelated pairs, thus effectively handling unequal class proportions and enhancing representation stability. The averaging-based pooling technique was implemented in this research to strengthen the generalization ability of causal tuple embeddings. In related research, Schrader et al. [15] designed a multi-layer sequence tagging approach strengthened with focused data augmentation. Their methodology handles cause–effect spans as structured sequences instead of individual tokens, thus improving boundary detection and reliability under noisy or limited training conditions. Alternative approaches have emphasized on effectiveness and flexibility. Burdisso et al. [16] proposed a *prompting-based few-shot learning* method that makes use of the hidden causal knowledge of large-scale pretrained language models using task-specific prompting. Their approach notably lowers dependency on labeled data while preserving strong evaluation performance. Similarly, Adibhatla and Shrivastava [17] introduced an *adapter-based fine-tuning* method, integrating lightweight domain-specific layers into pretrained transformer models. Their method showed pronounced flexibility over multiple domains as well as computational efficiency, hence making it suitable for resource-constrained situations. In summary, these contributions highlight two dominant patterns: (1) discriminative learning using organized training goals, such as contrastive learning and sequence tagging, which obtain high performance given sufficient supervision; and (2) parameter-efficient or few-shot learning approaches, including prompting and adapters, which focus on label efficiency and generalization across domains. Despite these advancements, multiple issues remain unaddressed. Hidden causal links and annotation ambiguity restrict current models, notably when causal cues depend on context or domain knowledge. Stability to domain changes and temporal variations is only partially handled, while model transparency at the level of span interpretation and consistency in evaluation across span, relation, and directionality metrics remains insufficiently explored.

2.2. Uncertainty Estimation and Ensemble Approaches in NLP and Causal Inference

Uncertainty estimation has emerged as a critical area in the field of natural language processing (NLP), reflecting the need for models to be more robust that are not only accurate but also reliable, interpretable, and calibrated in decision-making. Recent research as shown in Table 1 spans surveys, methodological innovations, and empirical evaluations, exploring how uncertainty manifests in data, models, and outputs, and how it can be quantified and leveraged for robustness, out-of-distribution (OOD) detection, calibration, and user trust.

At the intersection of causal inference and uncertainty quantification, recent research focuses on a broader span of theoretical foundations, Bayesian modeling, robust design-stage inference, and uncertainty-aware effect estimation. For instance, the quantification of causal emergence formalizes how macro-scale representations can exceed micro-scale

models through gains in effective information (*EI*), linking a single uncertainty parameter to changes in determinism and degeneracy [18]. Bayesian Interventional Mean Processes embed interventional distributions into a reproducing kernel Hilbert space (RKHS) with Gaussian process priors, propagating both aleatoric and epistemic uncertainty for improved causal Bayesian optimization under multimodal and covariate shifted settings [19]. Variational causal networks enable tractable ELBO optimization over DAG adjacency matrices with acyclicity-regularized priors, producing multimodal posteriors and outperforming factorized baselines in low-dimensional scenarios [20]. Robust design stage analyses formalize McNemar and z-tests based on integer programs to expose sensitivity to match assignments [21]. In time-series causal discovery, de-biased Lasso Granger estimators provide valid asymptotics under dependence and control false discovery rates while revealing real-world temporal relations [22]. Finally, uncertainty-aware neural CATE estimators leverage Bayesian parameter uncertainty and negative sampling to handle non-overlap as covariate shift, enabling informed deferral policies and reducing PEHE on datasets such as IHDP, ACIC, and CEMNIST [23].

Table 1. Summary of uncertainty-aware causal detection: contributions, uncertainty quantification (how uncertainty is measured or modeled), and qualification (how uncertainty is used for decisions, controls, or claims).

Paper	Contribution	Uncertainty Quantification	Uncertainty Qualification
[18]	Conditions for causal emergence via <i>EI</i> with determinism/degeneracy and threshold criteria	TPM-based determinism and degeneracy; CQE linking uncertainty parameter to <i>EI</i> gains	Use thresholds to justify/qualify coarse-graining and macro-level modeling claims
[19]	RKHS interventional embeddings with GP priors for $\mathbb{E}[T \mid do(X)]$ across fused graphs	Bayesian CMEs; GP mean/cov for interventional functionals; epistemic + aleatoric decomposition	Calibrated acquisition and decision-making in causal BO; uncertainty-aware fusion across sources
[20]	Autoregressive variational posterior over DAGs with acyclicity prior	ELBO-based approximate posterior over DAGs; multimodal structural uncertainty; metrics like Hellinger/ESHD	Structure-aware planning/experimentation under posterior mass over graphs; report credible structures
[21]	Robust hypothesis testing over feasible matches in matched designs	Worst/best-case <i>p</i> -value bounds via ILP over “good” match sets	Qualification of significance claims by reporting robust intervals; detect design fragility
[22]	De-biased Lasso–Granger with dependence-aware FDR control	Asymptotic variance for de-biased estimates; FDP/FDR control under dependence	Maintain valid discoveries (power vs. Type I) with FDR procedures tailored to dependence
[23]	Uncertainty-aware neural CATE with rejection/deferral policies	Epistemic variance via MC-Dropout/BNN; shift-aware diagnostics; PEHE/error vs. rejection curves	Reject unreliable recommendations under non-overlap/shift; safer individualized policies

Ensemble learning is commonly utilized to handle model uncertainty and inconsistency across multiple datasets. Traditional ensembles enhance generalization performance by merging predictions but often apply equal weights to all classifiers. Current studies address these issues by using entropy- or uncertainty-based weighting, which dynamically allocates trust based on model confidence. For example, Atmaja et al. (2024) introduced an entropy-based ensemble for speech classification that had better performance than single models and conventional ensemble schemes [24]. Ruan et al. (2022) put forward weighted-ensemble techniques in self-supervised learning (SSL), showing improved few-shot performance on ImageNet-1K benchmark [25]. Earlier studies by Sekhar et al. (2023) and Wang et al. (2017) used maximum entropy or information-entropy techniques to assign weights, obtaining strong adaptation in data streams and high performance across multiple datasets [26,27]. In the field of pattern recognition, Al-Maadeed et al. (2022) proved that the “entropy voting” of CNN ensembles exceeds majority-vote schemes in texture classification tasks [28].

Extending these insights, KL (Kullback–Leibler) divergence-based ensemble methods further lower predictive uncertainty and improve epistemic certainty. Malinin and

Gales (2019) proposed prior networks optimized using inverse KL targets to differentiate uncertainty types and strengthen out-of-distribution detection and selective prediction accuracy [29]. Sensoy et al. (2018) introduced an evidence-based deep learning framework that enables uncertainty-aware decision-making by capturing belief and predictive certainty [30]. Kumar et al. (2023) used KL-based regularization techniques for feature normalization in limited-data tasks, improving generalization to OOD and predictive calibration without directly changing output distributions [31]. Lee et al. (2018) leveraged KL divergence during training to promote high-entropy outputs on OOD inputs, improving calibration confidence [32]. Overall, these approaches show that KL divergence-based weighting enables the per-instance adjustment of model contributions, improving predictive confidence, reliability, and calibration, particularly in data-scarce or domain-shifted scenarios. Combining divergence-based objectives with adaptive ensemble methods allows models to better represent epistemic uncertainty while decreasing overall predictive uncertainty, aligning with current research directions in uncertainty-aware NLP and causal inference modeling.

2.3. Literature Review on Ensemble Methods

Ensemble-based strategies are widely employed across various domains, with significant use in NLP, especially for text classification scenarios. In the following subsection, we analyze significant studies that have aided to the creation and deployment of ensemble techniques in this research area. Mohammed and Kora [33] offered an in-depth review of ensemble deep learning techniques in NLP. They discussed several ensemble strategies, such as boosting, bagging, stacking, and negative correlation learning, and showed their potential to improve a model's overall performance. They further examined issues such as higher computational complexity and the risk of overfitting, providing scope for future research. Azim et al. [34] introduced a stacked ensemble approach that integrates several ML algorithms in order to enhance sentiment analysis on the IMDB review dataset. Their framework delivers superior performance, particularly for Urdu-linguistic data, by efficiently encoding contextual patterns and increasing classification accuracy. Kavitha and Prabhavathy [35] developed a hybrid framework that combines ensemble techniques with DL for text-based classification task. Their experimental evaluations examined combinations like Random Forest with Logistic Regression, showing enhanced classification accuracy over both textual and non-textual datasets. A recent survey [36] investigated ensemble methods employed in large language models (LLMs) for generating text and code. The authors discussed approaches such as weight fusion, knowledge integration, and mixture-of-experts architectures, showing how they resolve limitations of single models and increase generative scores. Ibrahim [37] carried out a comparative analysis of ensemble learning approaches versus traditional supervised ML methods for text classification tasks. Utilizing multiple datasets, the research evaluated the relative strengths and vulnerabilities of various ensemble strategies, providing insights into their real-world advantages and weakness.

While current studies illustrate the beneficial impact of ensemble methods in strengthening NLP classification performance, most prior works emphasize primarily on accuracy gains without explicitly tackling model trustworthiness, prediction uncertainty, or confidence calibration. This gap is critical in causal text detection, where misclassified causal relations can lead to incorrect reasoning in downstream tasks. Moreover, the discussed approaches typically rely on uniform or fixed ensemble strategies, ignoring that different base models may show varying confidence levels across instances. Hence, it is essential for an ensemble architecture that not only combines diverse embeddings and classifiers but also includes uncertainty estimates to dynamically adjust the model contributions.

Motivated by these limitations, the proposed work introduces an uncertainty-weighted ensemble architecture that leverages entropy and KL divergence to improve robustness, transparency, and predictive confidence in detecting causal relations.

3. Methodology

The proposed methodology as shown in Figure 1 in this work aims to overcome the issues in detecting causal relations, focusing on better classification and lowering uncertainty. The complete system combines various text embedding models with multiple classifiers and then applies uncertainty-based weighting to build a stronger ensemble classifier.

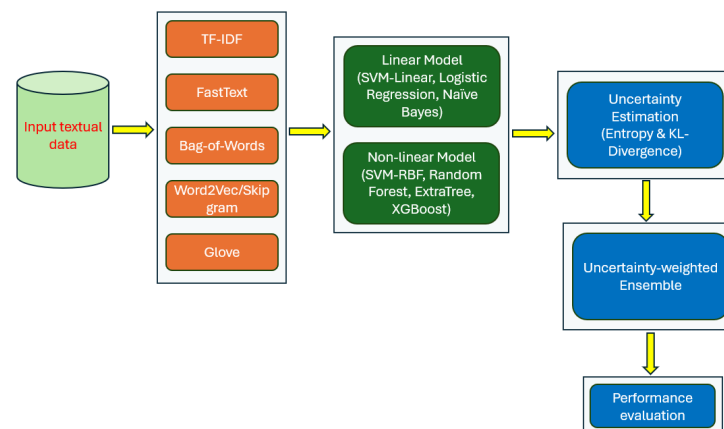


Figure 1. Pipeline of the proposed work to visualize the workflow.

3.1. ML Word Embeddings

To capture both lexical and distributional characteristics of causal textual data, we make use of a range of several embedding methods. TF-IDF offers a strong lexical foundation by measuring term importance within the corpus [38]. Bag-of-Words (BoW) further contributes sparse frequency-based features [39]. To combine distributional information, we use neural word embeddings produced using the Word2Vec model, namely CBOW and Skipgram, which create vector representations by predicting neighboring words [40]. These neural embeddings are often viewed as an implicit factorization of the Pointwise Mutual Information (PMI) matrix, thus extracting statistical co-occurrence in text. FastText which is developed by Facebook, expands the Word2Vec-Skipgram model by modeling each word as a set of character n-grams, allowing the effective handling of rare words, morphological changes, and unseen words [41]. In comparison, GloVe offers global co-occurrence embeddings by decomposing word-context statistics [42]. Overall, these embeddings cover a wide range from shallow lexical patterns (TF-IDF and BoW) to richer contextual co-occurrence relationships (Word2Vec, FastText, and GloVe).

3.2. Classifiers

To examine the predictive performance of various embedding methods, both linear and nonlinear classifiers were utilized. For linear models, we considered Support Vector Machine (SVM) with a linear kernel [43], Logistic Regression [44], and Naïve Bayes [45], which are effective in modeling linear decision boundaries and providing interpretable baselines. To account for nonlinear relationships within the data, we integrated radial basis function (RBF) kernel SVMs [46], as well as tree-based ensemble methods, including Random Forest [47], Extra Trees [48], and XGBoost [49]. The inclusion of these diverse classifiers allows the analysis to learn both linear and nonlinear relationships in causal text representations, thus ensuring a more stable and generalizable performance estimation across different embedding spaces.

3.3. Uncertainty-Aware Ensemble via Entropy-Based Weighted Ensemble

Instead of solely using the raw probability scores, this research architecture quantifies the classifier's uncertainty using Shannon entropy over the output probability distribution of each classifier. Entropy represents an interpretable metric of predictive confidence, where lower values indicate the model's certainty, and higher values indicate the model's uncertainty. For a given input x , entropy is defined as

$$H(x) = - \sum_{c=1}^C p(y = c | x) \log p(y = c | x) \quad (1)$$

where C is the number of classes. A low entropy shows confident predictions, while high entropy indicates uncertainty in prediction. The average entropy over the samples for each classifier offers a measure of uncertainty, directly aiding in the weighting strategy in the ensemble.

Entropy-Guided Ensemble Fusion

The final decision is obtained using a weighted ensemble approach, where predictions from individual classifiers are combined based on entropy-guided weights. The low-entropy classifiers are assigned higher weights due to their better predictive confidence level; as a result, the higher entropy classifiers with low confidence contributes minimal to the final decision process. The dynamic weighting scheme facilitates the ensemble weighting classifier to utilize the strength of the high-confident classifier over less-confident classifiers. The assigned weights to each classifier i is negatively correlated with its uncertainty quantified through entropy estimation and is defined as

$$w_i = \frac{e^{-H_i}}{\sum_{j=1}^M e^{-H_j}} \quad (2)$$

where H_i is the mean entropy of classifier i and M is the total number of classifiers. This makes sure that high-confidence (low-entropy) classifiers have greater influence in ensemble decision-making. The final ensemble prediction is evaluated as a weighted aggregation of classifier outputs:

$$P(y | x) = \sum_{i=1}^M w_i p_i(y | x) \quad (3)$$

The predicted class is obtained by selecting the label with the highest probability in $P(y | x)$. This framework minimizes classifiers' overconfident predictions. This adaptively down-weights uncertain classifiers and reduces the overall predictive uncertainty, thereby improving robustness.

3.4. Uncertainty-Aware Ensemble via KL Divergence

In this research, we discuss the uncertainty-aware ensemble framework where classifiers' weights are derived from the Kullback–Leibler (KL) divergence between model predictions and true label probabilities.

3.4.1. KL Divergence for Uncertainty Estimation

For a single instance with true class label y , we define the true label distribution as a one-hot vector $\mathbf{p}_{\text{true}}$ and the predicted probability distribution from a base model as $\mathbf{p}_{\text{pred}}$. The KL divergence is computed as

$$D_{\text{KL}}(\mathbf{p}_{\text{true}} \parallel \mathbf{p}_{\text{pred}}) = \sum_{i=1}^C \mathbf{p}_{\text{true}}(i) \log \frac{\mathbf{p}_{\text{true}}(i)}{\mathbf{p}_{\text{pred}}(i)} = -\log \mathbf{p}_{\text{pred}}(y), \quad (4)$$

where C is the number of classes. Lower KL indicates higher confidence and correct predictions, whereas higher KL reflects uncertain or incorrect predictions. The mean KL divergence across all test instances serves as an aggregate uncertainty measure for each base model:

$$\text{Mean KL} = \frac{1}{N} \sum_{j=1}^N D_{\text{KL}}(\mathbf{p}_{\text{true}}^{(j)} \parallel \mathbf{p}_{\text{pred}}^{(j)}), \quad (5)$$

where N is the number of test samples.

3.4.2. Inverse-KL Weighting for Ensemble

To construct the ensemble, we assign weights to each base model inversely proportional to the square root of its mean KL divergence:

$$w_m = \frac{1 / \sqrt{\max(\text{Mean KL}_m, \epsilon)}}{\sum_k 1 / \sqrt{\max(\text{Mean KL}_k, \epsilon)}}, \quad (6)$$

where w_m is the normalized weight of model m , and ϵ is a small floor value (e.g., 10^{-3}) to avoid division by zero.

3.4.3. Ensemble Prediction

The ensemble predicted probabilities are obtained by a weighted average of the base classifiers predictions:

$$\mathbf{p}_{\text{ensemble}} = \sum_m w_m \mathbf{p}_m, \quad (7)$$

where $\mathbf{p}_m$ denotes the predicted probability vector of model m . The final class label is assigned as

$$\hat{y}_{\text{ensemble}} = \arg \max_i \mathbf{p}_{\text{ensemble}}(i). \quad (8)$$

4. Experiments and Results

To experimentally validate the proposed uncertainty-weighted ensemble model for causal relation detection in textual data, we designed a set of experiments combining diverse embedding techniques with multiple classifiers. The experimental setup aimed to analyze not only the independent performance of various embeddings and classifiers but also the effectiveness of incorporating uncertainty estimates into the ensemble decision-making mechanism. All experiments were conducted on a system featuring an Intel Core i5 11300H CPU (3.10 GHz), 16 GB RAM, running 64-bit Windows 10. The extraction of textual features employing ML embeddings, such as TF-IDF, Bag-of-Words (BoW), Word2Vec-CBOW, Skip-Gram, GloVe, and FastText, enables representation of both traditional and dense semantic representations. These embeddings are employed to train a diverse range of classifiers, both linear and nonlinear classifiers, for developing an uncertainty-aware framework suitable for both linear and nonlinear learners. For each classifier, predictive uncertainty was measured using “entropy-based” and “KL divergence-based” measures derived from the output probability distributions. The uncertainty scores were subsequently leveraged to assign adaptive weights in the ensemble, allowing more confident models to dominate the final prediction. All results were evaluated using a 10-fold cross-validation strategy to ensure robustness and generalizability. In this section, we present the results of these experiments, compare the performance of baseline models with our uncertainty-weighted ensemble, and analyze the improvements in both accuracy and reliability achieved by the proposed approach.

4.1. Data Descriptions

4.1.1. Dataset—1 (Causal News Corpus (CNC))

The Causal News Corpus (CNC) dataset [50] introduced by Tan et al. (2022) is a richly annotated dataset designed for causal relation detection in news text, comprising 3559 event-driven sentences each labeled for the presence of cause-and-effect relation in the corpus. This dataset is a part of the Codalab shared task containing train, dev and test data with 3248 (combining both train and dev) and 311 (test) entries. The dataset had a binary class representation as 1 for causal and 0 for non-causal representations. The class-wise distribution of each class is shown in Figure 2. This dataset contains both implicit and explicit forms of causal contents. The average number of tokens, minimum and maximum tokens, and the vocabulary size for each class in Dataset 1 are shown in Table 2.

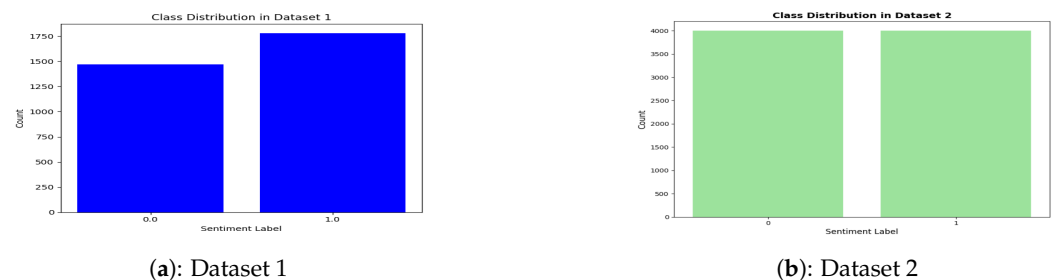


Figure 2. Class-wise distributions for all datasets.

Table 2. Class-wise token statistics for Dataset 1 and Dataset 2.

Dataset	Class	AvgTokens	Min	Max	VocabSize
Dataset 1	0	23.72	4	116	6403
	1	31.33	6	133	8097
Dataset 2	0	134.39	5	2891	28,459
	1	134.70	6	1321	30,778

4.1.2. Dataset—2 (CausalLM–Adjective Group)

The CausalLM dataset [51] Feder, 2020, was introduced to facilitate causal language modeling and explanation in natural language. It was designed to enable counterfactual text generation and causal relation analysis, going beyond surface level correlation by focusing on linguistic constructs that imply causality. This dataset is taken from the Kaggle repository from the CausalLM–Adjective group. The dataset is split into test, train and dev sets. A balanced dataset with an equal number of representing causal and non-causal (Figure 2) entities having both explicit and implicit causality in the corpus. Table 2 shows the average number of tokens, as well as the minimum and maximum token counts, and the vocabulary size for each class in Dataset 2.

4.2. Hyperparameter Tuning

To evaluate the best performance of the models and ensure their generalization ability, we performed hyperparameter tuning separately for linear and nonlinear classifiers across all textual embeddings. For the linear classifiers—SVM with a linear kernel, Logistic Regression, and Naïve Bayes—the main optimization variables are the regularization parameter C for SVM and Logistic Regression, and the smoothing parameter α for Naïve Bayes. Importantly, these regularization parameters are independent across classifiers since each model optimizes a different objective function and employs a distinct regularization scheme; therefore, no coupling or shared regularization was applied between models. For the nonlinear classifiers, such as SVM with an RBF kernel, Random Forest, Extra Trees,

and XGBoost, the optimization focused on kernel parameters (C , γ , degree) and ensemble-related hyperparameters including the number of estimators, maximum depth, learning rate, and subsampling ratio. The final optimized parameter settings are shown in Table 3 for linear classifiers and Table 4 for nonlinear classifiers.

Table 3. Optimal hyperparameter configurations for linear classifiers across all textual embeddings. Each classifier's hyperparameters are tuned independently, reflecting their distinct regularization mechanisms.

Embedding	Classifier	Best Hyperparameters
TF-IDF	SVM (Linear)	$C = 1.0$
	Logistic Regression	$C = 0.1$, penalty = L_2
	Naïve Bayes	$\alpha = 0.01$
BoW	SVM (Linear)	$C = 0.1$
	Logistic Regression	$C = 1.0$, penalty = L_2
	Naïve Bayes	$\alpha = 0.1$
Word2Vec-CBOW	SVM (Linear)	$C = 10$
	Logistic Regression	$C = 1.0$, penalty = L_2
	Naïve Bayes	$\alpha = 0.05$
Word2Vec-Skipgram	SVM (Linear)	$C = 1.0$
	Logistic Regression	$C = 0.1$, penalty = L_2
	Naïve Bayes	$\alpha = 0.01$
GloVe	SVM (Linear)	$C = 10$
	Logistic Regression	$C = 1.0$, penalty = L_2
	Naïve Bayes	$\alpha = 0.1$
FastText	SVM (Linear)	$C = 0.1$
	Logistic Regression	$C = 1.0$, penalty = L_2
	Naïve Bayes	$\alpha = 0.01$

Table 4. Optimal hyperparameter configurations for Non-linear classifiers across all textual embeddings. Each classifier's hyperparameters are tuned independently, reflecting their distinct regularization mechanisms.

Embedding	Classifier	Best Hyperparameter
TF-IDF	SVM (RBF)	$C = 10$, $\gamma = 0.01$
	Random Forest	n_estimators = 200, max_depth = 20
	Extra Trees	n_estimators = 200, max_depth = None
	XGBoost	n_estimators = 500, learning_rate = 0.05, max_depth = 5, $\lambda = 1$, $\alpha = 0.1$
BoW	SVM (RBF)	$C = 100$, $\gamma = 0.001$
	Random Forest	n_estimators = 100, max_depth = 50
	Extra Trees	n_estimators = 200, max_depth = None
	XGBoost	n_estimators = 500, learning_rate = 0.1, max_depth = 3, $\lambda = 0$, $\alpha = 0.1$
Word2Vec-CBOW	SVM (RBF)	$C = 1.0$, $\gamma = 0.01$
	Random Forest	n_estimators = 200, max_depth = 20
	Extra Trees	n_estimators = 500, max_depth = None
	XGBoost	n_estimators = 200, learning_rate = 0.1, max_depth = 3, $\lambda = 1$, $\alpha = 0$
Word2Vec-Skipgram	SVM (RBF)	$C = 10$, $\gamma = 0.001$
	Random Forest	n_estimators = 200, max_depth = 20
	Extra Trees	n_estimators = 500, max_depth = 50
	XGBoost	n_estimators = 500, learning_rate = 0.05, max_depth = 5, $\lambda = 1$, $\alpha = 0.1$
GloVe	SVM (RBF)	$C = 100$, $\gamma = 0.01$
	Random Forest	n_estimators = 200, max_depth = 20
	Extra Trees	n_estimators = 200, max_depth = None
	XGBoost	n_estimators = 500, learning_rate = 0.1, max_depth = 3, $\lambda = 0.1$, $\alpha = 0.1$
FastText	SVM (RBF)	$C = 10$, $\gamma = 0.01$
	Random Forest	n_estimators = 200, max_depth = 20
	Extra Trees	n_estimators = 500, max_depth = 50
	XGBoost	n_estimators = 200, learning_rate = 0.1, max_depth = 5, $\lambda = 1$, $\alpha = 0$

4.3. Results of ML-Embedding on Classifiers

In this research, we discuss the effect of both the linear and nonlinear approaches in the classification of causal text.

4.3.1. Linear Classifiers

As shown in Table 5, the linear models, namely SVM Linear, Logistic Regression, and Naive Bayes, demonstrate moderate predictive capability on both Dataset 1 (News Corpus) and Dataset 2 (CausalLM–Adjective group). It has been observed that even combining with high-quality text embeddings, these classifiers rarely exceed 90%, indicating that the nonlinear decision boundaries are insufficient to fully capture the complex, nonlinear structures underlying the causal relationships in the datasets. It is observed on Dataset 1 that the TF-IDF word embedding techniques with SVM Linear got the best results in linear classifier with accuracy, precision, F1-score and recall values of 91.33%, 91.31%, 91.30% and 91.33% respectively. On Dataset 2, the Logistic Regression model outperforms the other linear classifiers, achieving accuracy, precision, F1-score and recall values of 92.0%, 91.8%, 92.0% and 92.2%, respectively.

Table 5. Performance comparison of embeddings with linear classifiers (SVM-Linear, Logistic Regression and Naive Bayes) on Dataset 1 (News Corpus) and Dataset 2 (CausalLM–Adjective group).

Metric	Embedding	Dataset 1			Dataset 2		
		SVM-Linear	Logistic Regression	Naive Bayes	SVM-Linear	Logistic Regression	Naive Bayes
Accuracy	GloVe	66.25	64.70	57.40	66.5	67.0	70.2
	Skip-gram	67.18	69.66	59.35	75.5	78.5	72.5
	FastText	66.56	71.82	60.12	70.0	73.5	70.0
	Word2Vec-CBOW	75.23	84.21	63.25	75.5	78.5	71.8
	BoW	83.32	78.94	72.10	89.5	83.0	73.5
	TF-IDF	91.33	86.45	77.52	91.0	92.0	75.4
Precision	GloVe	66.08	64.50	57.20	66.3	66.8	70.0
	Skip-gram	67.06	69.55	59.10	75.2	78.2	72.2
	FastText	68.04	72.48	60.42	69.8	73.0	69.8
	Word2Vec-CBOW	75.22	84.30	63.10	75.2	78.0	71.5
	BoW	83.10	85.66	72.30	89.0	82.5	73.2
	TF-IDF	91.31	86.31	77.18	90.8	91.8	75.1
Recall	GloVe	66.25	64.70	57.30	66.8	67.2	70.5
	Skip-gram	67.18	69.66	59.20	75.8	78.7	72.8
	FastText	66.56	71.82	60.00	70.5	73.5	70.3
	Word2Vec-CBOW	75.23	84.21	63.05	75.8	78.7	72.0
	BoW	83.32	78.94	72.00	89.0	82.8	73.8
	TF-IDF	91.33	86.45	77.25	91.2	92.2	75.8
F1-score	GloVe	65.82	64.10	57.25	66.5	66.8	70.2
	Skip-gram	67.08	69.39	59.15	75.5	78.5	72.5
	FastText	64.49	71.02	60.05	70.0	73.5	70.0
	Word2Vec-CBOW	75.04	84.11	63.12	75.5	78.5	71.8
	BoW	83.01	78.46	72.05	89.5	83.0	73.5
	TF-IDF	91.30	86.23	77.33	91.0	92.0	75.4

Note: Bold values indicate the best-performing embedding–classifier combination for each evaluation metric.

4.3.2. Nonlinear Classifiers

In contrast with the linear classifiers, the nonlinear classifiers—namely XGBoost, SVM-RBF, Random Forest, and Extra Trees—are capable of modeling complex, nonlinear decision boundaries. It is observed that the kernel and tree-based approaches achieve a higher performance metric compared to the linear classifiers. On Dataset 1, the TF-IDF embedding with the SVM-RBF classifier achieved an accuracy of 94.53% with a precision, recall and F1-score values of 94.50%, 94.42% and 94.61%, respectively. Similarly, for Dataset 2, the same approach acquired the best result with accuracy, precision, F1-score and recall

of 94.5%, 94.4%, 94.5% and 94.6% respectively. Detailed results observed on different nonlinear classifiers are shown in Table 6.

Table 6. Performance comparison of embeddings with nonlinear classifiers (SVM-RBF, Random Forest, Extra Trees, and XGBoost) on Dataset 1 (News Corpus) and Dataset 2 (CausalLM–Adjective group).

Metric	Embedding	Dataset 1				Dataset 2			
		XGBoost	RF	SVM-RBF	Extra Trees	SVM-RBF	RF	Extra Trees	XGBoost
Accuracy	GloVe	65.32	67.28	71.42	65.30	94.2	89.3	89.8	93.1
	Skip-gram	65.89	64.30	86.82	78.52	93.2	90.5	89.9	91.0
	FastText	64.65	73.62	83.52	61.82	91.7	88.9	89.2	90.4
	Word2Vec-CBOW	69.59	70.82	94.42	81.52	92.8	89.7	88.6	91.5
	BoW	93.49	79.56	93.42	71.82	93.8	91.5	90.7	92.9
	TF-IDF	94.21	81.82	94.53	93.32	94.5	91.2	90.0	93.6
Precision	GloVe	65.21	67.16	72.18	65.28	94.0	89.9	88.7	93.1
	Skip-gram	65.60	64.28	86.75	78.15	93.1	90.4	90.1	91.2
	FastText	64.50	73.30	83.15	62.48	91.5	89.6	89.8	90.2
	Word2Vec-CBOW	69.41	71.48	94.50	82.48	92.6	90.1	89.4	91.8
	BoW	93.52	79.42	93.50	72.48	93.6	90.2	90.5	92.8
	TF-IDF	94.23	82.48	94.50	93.10	94.4	90.8	90.2	93.7
Recall	GloVe	65.32	67.48	71.82	65.40	94.1	89.4	89.0	93.0
	Skip-gram	65.89	64.40	86.32	78.22	93.0	90.7	89.5	91.1
	FastText	64.65	73.26	83.22	61.82	91.6	89.0	89.3	90.0
	Word2Vec-CBOW	69.59	70.82	94.42	81.82	92.9	88.8	88.5	91.3
	BoW	93.49	79.36	93.42	71.82	93.5	91.2	90.8	92.7
	TF-IDF	94.21	81.82	94.42	93.32	94.6	91.0	89.7	93.5
F1-score	GloVe	65.08	67.28	71.32	65.35	94.2	89.6	88.9	93.0
	Skip-gram	65.51	64.35	86.41	78.31	93.1	90.3	89.8	91.2
	FastText	64.32	73.01	83.31	61.02	91.7	89.1	89.0	90.3
	Word2Vec-CBOW	69.32	70.02	94.61	81.32	92.8	89.5	89.1	91.5
	BoW	93.50	79.53	93.61	71.02	93.7	91.0	90.2	92.8
	TF-IDF	95.01	81.02	94.61	93.01	94.5	90.9	89.9	93.6

Note: Bold values indicate the best-performing embedding–classifier combination for each evaluation metric.

4.4. Entropy-Based Uncertainty Estimation

The Shannon entropy values reflect the degree of uncertainty in the predictions of the base classifiers. A lower entropy indicates higher confidence in predictions, while a higher entropy suggests greater uncertainty. The entropy observed in each dataset on both linear and nonlinear classifiers is shown in Table 7 and Table 8, respectively. It is noticed that among linear classifiers, SVM-Linear exhibits lower entropy compared to Naïve Bayes and Logistic Regression. Similarly, in nonlinear classifiers, the Extra Tree model has the lowest entropy, followed by the SVM-RBF, Random Forest and XGBoost. This emphasizes the importance of entropy-based weighting in ensembles, where models with lower entropy contribute more reliable predictions.

Table 7. Entropy values of linear classifiers (SVM-Linear, Logistic Regression, and Multinomial Naive Bayes) across six embedding techniques for Dataset 1 (News Corpus) and Dataset 2 (CausalLM–Adjective group).

Embedding	Dataset 1			Dataset 2		
	SVM-Linear	Logistic Regression	Naive Bayes	SVM-Linear	Logistic Regression	Naive Bayes
GloVe	0.61	0.71	0.65	0.57	0.69	0.63
Skip-gram	0.60	0.69	0.64	0.55	0.67	0.61
FastText	0.58	0.68	0.63	0.53	0.65	0.60
Word2Vec-CBOW	0.59	0.70	0.64	0.54	0.67	0.62
BoW	0.45	0.62	0.55	0.36	0.56	0.52
TF-IDF	0.48	0.63	0.58	0.36	0.56	0.55

Table 8. Entropy values of nonlinear classifiers (SVM-RBF, Random Forest, Extra Trees, and XGBoost) on Dataset 1 and Dataset 2 for all embeddings.

Embedding	Dataset 1				Dataset 2			
	SVM-RBF	Random Forest	Extra Trees	XGBoost	SVM-RBF	Random Forest	Extra Trees	XGBoost
GloVe	0.60	0.78	0.83	0.72	0.28	0.36	0.34	0.30
Skip-gram	0.41	0.80	0.60	0.55	0.30	0.34	0.32	0.29
FastText	0.46	0.70	0.68	0.60	0.33	0.39	0.36	0.32
Word2Vec-CBOW	0.28	0.65	0.54	0.48	0.31	0.37	0.35	0.31
BoW	0.26	0.50	0.52	0.40	0.27	0.33	0.30	0.28
TF-IDF	0.26	0.41	0.14	0.46	0.19	0.41	0.13	0.41

Entropy-Based Weighted-Ensemble Classifiers

In this research, the entropy-based weighting is assigned to the classifiers to develop an uncertainty-aware machine learning model to have more certain and accurate predictions. The model with lower entropy, suggested to have a better certain prediction, is assigned to have higher weights and vice versa. In this research, individual performance on both the linear and nonlinear approaches, along with entropy estimation and entropy-based weighting, is given to each category. The entropy-based weighting of the linear classifiers is shown in Table 9, in which the SVM-Linear model is assigned more weights, followed by the Naive Bayes and Logistic Regression. Table 10 shows the entropy-based weights assigned to the nonlinear classifiers.

Table 9. Ensemble weights of linear base models (SVM-Linear, Naive Bayes, and Logistic Regression) across all embeddings for Dataset 1 and Dataset 2.

Classifier	Dataset 1						Dataset 2					
	TF-IDF	BoW	Word2Vec-CBOW	FastText	Skip-Gram	GloVe	TF-IDF	BoW	Word2Vec-CBOW	FastText	Skip-Gram	GloVe
SVM-Linear	0.398	0.391	0.382	0.368	0.362	0.355	0.459	0.405	0.376	0.360	0.370	0.348
Naive Bayes	0.306	0.310	0.315	0.318	0.322	0.330	0.311	0.305	0.320	0.323	0.320	0.335
Logistic Regression	0.296	0.299	0.303	0.314	0.316	0.315	0.230	0.290	0.304	0.317	0.310	0.317

Table 10. Ensemble weights of nonlinear classifiers (SVM-RBF, Random Forest, Extra Trees, and XGBoost) across all embeddings for Dataset 1 and Dataset 2.

Classifier	Dataset 1						Dataset 2					
	TF-IDF	BoW	Word2Vec-CBOW	FastText	Skip-Gram	GloVe	TF-IDF	BoW	Word2Vec-CBOW	FastText	Skip-Gram	GloVe
SVM-RBF	0.342	0.318	0.276	0.265	0.248	0.223	0.331	0.309	0.272	0.259	0.246	0.221
Random Forest	0.192	0.221	0.237	0.248	0.266	0.281	0.181	0.198	0.228	0.234	0.257	0.273
Extra Trees	0.408	0.377	0.358	0.352	0.336	0.335	0.438	0.398	0.382	0.368	0.351	0.352
XGBoost	0.058	0.084	0.129	0.135	0.150	0.161	0.050	0.095	0.118	0.139	0.146	0.154

The ensemble model developed, assigning the weights obtained from the entropy-based approach, performed better compared to the individual performance of each classifier (linear and nonlinear). Table 11 clearly shows the performance of the ensemble approach on both datasets using Linear classifiers. The best model achieved an accuracy value of 92.74% with precision and recall values of 91.82% and 91.96%. The value of F1-score is 91.88%. Uncertainty estimations reveal better performance with lower entropy values of around 0.4237 and with a prediction confidence value of 0.7989 on Dataset 1. Similarly, on Dataset 2, in the entropy-based weighted ensemble approach, the model achieved an accuracy of 94.28% and an F1-score value of 94.01%, and recall and precision values of 94.05% and 93.97%, respectively. The prediction confidence achieved on Dataset 2 using this approach is 0.7732. The variation ratio observed in both Dataset 1 and Dataset 2 was 0.0943 and 0.00728, respectively. Table 12 shows the performance using the nonlinear classifiers on both datasets. On Dataset 1, the model's prediction confidence value achieved was around 0.7217 with a variation ratio of 0.0824; similarly, on Dataset 2, the model's prediction confidence was 0.7985 and a variation ratio of around 0.0672. The model achieved an accuracy value of 95.52% with precision and recall values of 94.80% and

94.92%. The F1-score value observed around 94.65% on Dataset 1. Similarly, on Dataset 2, in the entropy-based weighted ensemble approach, the model achieved an accuracy of 96.89% and an F1-score value of 96.87%, the recall and precision values of 96.85% and 96.92%, respectively. Overall, the results confirm that the proposed uncertainty-weighted ensemble model not only achieves better prediction accuracy but also ensures enhanced reliability and interpretability, where uncertainty calibration can be substantially improved.

Table 11. Entropy-based ensemble performance metrics of linear classifiers (SVM-Linear, Logistic Regression, and Naive Bayes) across six embeddings for Dataset 1 (News Corpus) and Dataset 2 (CausalLM–Adjective group), excluding ROC-AUC.

Embedding	Dataset 1								Dataset 2							
	Acc	Prec	Rec	F1	Entropy	Conf. Unc.	Pred. Conf.	Var. Ratio	Acc	Prec	Rec	F1	Entropy	Conf. Unc.	Pred. Conf.	Var. Ratio
TF-IDF	0.927	0.918	0.920	0.919	0.424	0.201	0.799	0.094	0.943	0.941	0.940	0.940	0.382	0.227	0.774	0.073
BoW	0.912	0.903	0.905	0.904	0.465	0.238	0.762	0.102	0.927	0.925	0.924	0.924	0.415	0.242	0.758	0.082
Word2Vec-CBOW	0.899	0.891	0.892	0.891	0.487	0.253	0.745	0.108	0.918	0.916	0.915	0.915	0.435	0.253	0.743	0.088
FastText	0.902	0.895	0.896	0.895	0.482	0.249	0.750	0.106	0.920	0.918	0.917	0.917	0.430	0.250	0.746	0.086
Skip-gram	0.900	0.892	0.894	0.893	0.485	0.251	0.748	0.107	0.919	0.917	0.916	0.916	0.432	0.251	0.745	0.087
GloVe	0.905	0.897	0.898	0.897	0.478	0.246	0.755	0.105	0.922	0.920	0.919	0.919	0.428	0.248	0.748	0.085

Note: Bold values denote the best-performing embedding configuration for each metric within the corresponding dataset.

Table 12. Ensemble performance metrics of nonlinear classifiers (SVM-RBF, Random Forest, Extra Trees, and XGBoost) for Dataset 1 (News Corpus) and Dataset 2 (CausalLM–Adjective group) across six embeddings, excluding ROC-AUC.

Embedding	Dataset 1								Dataset 2							
	Acc	Prec	Rec	F1	Entropy	Conf. Unc.	Pred. Conf.	Var. Ratio	Acc	Prec	Rec	F1	Entropy	Conf. Unc.	Pred. Conf.	Var. Ratio
TF-IDF	0.955	0.948	0.949	0.947	0.413	0.278	0.722	0.082	0.969	0.969	0.969	0.969	0.352	0.202	0.798	0.067
BoW	0.942	0.935	0.936	0.935	0.435	0.295	0.705	0.089	0.955	0.954	0.953	0.954	0.372	0.215	0.785	0.073
Word2Vec-CBOW	0.938	0.931	0.932	0.931	0.445	0.302	0.698	0.091	0.950	0.948	0.948	0.948	0.382	0.222	0.778	0.075
FastText	0.940	0.933	0.934	0.933	0.440	0.298	0.702	0.090	0.952	0.950	0.950	0.950	0.378	0.219	0.782	0.074
Skip-gram	0.939	0.932	0.933	0.932	0.442	0.300	0.700	0.090	0.951	0.949	0.949	0.949	0.380	0.221	0.780	0.074
GloVe	0.941	0.934	0.935	0.934	0.438	0.296	0.704	0.089	0.953	0.951	0.951	0.951	0.376	0.218	0.782	0.073

Note: Bold values denote the best-performing embedding configuration for each evaluation metric within the corresponding dataset.

4.5. KL Divergence-Based Uncertainty Estimation

To evaluate the model uncertainty across the multiple ML embeddings across both the linear and nonlinear classifiers, we integrate the Kullback–Leibler (KL) divergence approach, which evaluates the probability of the predicted class and the true one-hot labels. In this section, we discuss the detailed result analysis of the effect of KL divergence on both the linear and nonlinear classifiers and the effect on multiple embeddings. Along with the standard classification metrics: accuracy, precision, recall, F1-score, and LogLoss. Additionally, the mean KL divergence of the ensemble predictions provides a measure of overall confidence in the predictions.

4.5.1. KL Divergence-Based Uncertainty on Classifiers

From Table 13, the predictive uncertainty of the three linear classifiers—Logistic Regression, SVM-Linear, and Naïve Bayes—across the six embedding methods (TF-IDF, BoW, Word2Vec-CBOW, FastText, Skip-gram, and GloVe) can be observed. Uncertainty-based performance evaluation of each classifier was achieved using the LogLoss and Mean KL divergence, which collectively captured the calibration and consistency of the probabilities. It is observed from the table that the Linear SVM achieves the best calibration score with the lowest LogLoss on both the dataset and the least KL divergence, followed by the logistic and the naive bayes with the highest KL divergence score.

The LogLoss and KL Mean of the nonlinear classifiers on the six different embeddings on both datasets are shown in Table 14. Both these metrics jointly help in predictive accuracy and probabilistic calibration analysis. It is observed that the SVM RBF kernel yields the lowest LogLoss on Dataset 1 with a value of 0.195 for the Word2Vec-CBOW embedding technique and the lowest KL Mean value compared to the other nonlinear classifiers like ExtraTree, Random Forest and XgBoost. It is noticed that the Random Forest and Extra Tree classifiers show higher KL Mean and the least calibration compared to other classifiers, despite competitive LogLoss, showcasing their more distinct probability estimates. It is observed that the TF-IDF and BoW embedding got lower LogLoss and KL Mean compared to the dense embeddings, highlighting the robustness of sparse lexical features suitable for probabilistic modeling. It is also observed that Word2Vec-CBOW embedding achieved the best LogLoss for SVM-RBF on Dataset 1, highlighting the advantage of pairing dense semantic representation with flexible kernels.

Table 13. LogLoss and mean Kullback–Leibler (KL) divergence of linear classifier models across six embeddings for Dataset 1 (News Corpus) and Dataset 2 (CausalLM–Adjective group).

Classifiers	Metrics	Dataset 1						Dataset 2					
		TF-IDF	BoW	Word2Vec-CBOW	FastText	Skip-Gram	GloVe	TF-IDF	BoW	Word2Vec-CBOW	FastText	Skip-Gram	GloVe
Logistic Regression	LogLoss	0.458	0.472	0.485	0.482	0.483	0.480	0.399	0.415	0.428	0.425	0.426	0.423
	KL Mean	0.079	0.089	0.094	0.092	0.093	0.091	0.080	0.091	0.095	0.094	0.095	0.093
Linear SVM	LogLoss	0.315	0.328	0.340	0.338	0.339	0.336	0.270	0.285	0.298	0.296	0.297	0.294
	KL Mean	0.224	0.232	0.240	0.238	0.239	0.237	0.225	0.233	0.241	0.239	0.240	0.238
Naive Bayes	LogLoss	0.456	0.470	0.483	0.480	0.481	0.478	0.463	0.477	0.489	0.486	0.487	0.484
	KL Mean	0.097	0.108	0.114	0.112	0.113	0.111	0.098	0.109	0.115	0.113	0.114	0.112

Table 14. LogLoss and mean Kullback–Leibler (KL) divergence of nonlinear Classifier models across six embeddings for Dataset 1 (News Corpus) and Dataset 2 (CausalLM–Adjective group).

Classifiers	Metrics	Dataset 1						Dataset 2					
		TF-IDF	BoW	Word2Vec-CBOW	FastText	Skip-Gram	GloVe	TF-IDF	BoW	Word2Vec-CBOW	FastText	Skip-Gram	GloVe
SVM-RBF	LogLoss	0.242	0.263	0.195	0.221	0.208	0.230	0.310	0.287	0.301	0.298	0.295	0.312
	KL Mean	0.142	0.154	0.098	0.115	0.107	0.123	0.162	0.149	0.153	0.150	0.148	0.160
Random Forest	LogLoss	0.318	0.295	0.310	0.342	0.328	0.335	0.287	0.263	0.270	0.280	0.275	0.290
	KL Mean	0.182	0.171	0.176	0.190	0.185	0.188	0.159	0.145	0.148	0.153	0.150	0.160
Extra Trees	LogLoss	0.305	0.312	0.298	0.321	0.308	0.315	0.293	0.280	0.285	0.292	0.289	0.298
	KL Mean	0.175	0.179	0.168	0.182	0.174	0.178	0.160	0.152	0.155	0.160	0.158	0.164
XGBoost	LogLoss	0.228	0.242	0.235	0.240	0.238	0.245	0.270	0.258	0.265	0.268	0.266	0.272
	KL Mean	0.138	0.146	0.142	0.144	0.143	0.147	0.150	0.144	0.147	0.149	0.148	0.152

4.5.2. KL-Based Ensemble Weights for Classifiers

In this section, we discuss the KL-based ensemble weighting for uncertainty reduction in both linear and nonlinear classifiers. The higher weights are assigned to the models with the lower KL divergence by directly favoring the classifiers for better calibration probability. Table 15 presents the weights assigned to the linear classifiers on the multiple ML Embeddings on Dataset 1 and Dataset 2. It is observed that the Logistic Regression is assigned the highest weights reflecting its stronger calibration performance. Table 16 shows the weights assigned to the nonlinear classifiers. The SVM–RBF model receives the highest weight due to its higher confidence and better calibration across the six embedding methods and the two datasets.

An uncertainty-aware ensemble model in which the prediction of the classifiers, combining the inverse KL divergence weighting, is integrated for developing a more confident model for analysis. This approach of KL divergence integration in uncertainty aims to assign higher weights to the classifiers having lower mean predictive divergence, hence favoring models to produce better calibrated results. In this section, we discuss the KL divergence-based weighting results for both linear and nonlinear classifiers. Table 17 displays the performance of KL divergence-based ensemble on linear classifiers on both

Dataset 1 and Dataset 2. The TF-IDF word embeddings technique with the weighted ensemble approach got the highest compared to other embedding techniques on Dataset 1, with an accuracy value of 89.8% and an F1-score of 89.6% with a LogLoss of 0.430 and a KL Mean of around 0.097, showcasing that the sparse lexical representation yields the most robust probability estimates. Similarly, on Dataset 2, the TF-IDF model got an accuracy of around 74.5%, an F1-score of around 74.4, a LogLoss of 0.522 and a KL Mean of around 0.567. The KL divergence-based weighting on the nonlinear classifiers proved to have a better result and calibration compared with the entropy-based ensemble approach, with the highest F1-score value noticed on Dataset 1 to be around 95.6% and with a LogLoss of around 0.292, with a notably low KL mean divergence of around 0.072. The result of each embedding with the KL divergence-based weighted ensemble is shown in Table 18. On Dataset 2, 97.6% accuracy along with an F1-score of 97.2% with a LogLoss and a KL Mean of values 0.221 and 0.055 was observed for TF-IDF embeddings. The ensemble approach effectively utilizes the calibration advantages of SVM-RBF and XgBoost, resulting in better performance and confident prediction. The inverse KL-weighting integrated ensures that the classifiers with the highest confidence and calibration will be contributing to the ensemble model, thus having a strong, robust model with consistent F1-score, accuracy, and low KL divergence. Sparse embedding consistently having superior performance across both datasets showcases the advantages of sparse, frequency-based features for uncertainty-aware ensemble learning models.

The experimental evaluation shows that the **entropy-based weighting** method enhanced both calibration and prediction confidence for **linear classifiers**, achieving an accuracy of 94.3%, an F1-score of 94.0%, and an entropy value of 0.382, with an average confidence of 0.774. In contrast, the **KL divergence-based weighting** approach yielded superior outcomes with **nonlinear classifiers**, attaining an accuracy of 97.6% and an F1-score of 97.2%, while maintaining a low KL mean of 0.055 and a LogLoss of 0.221. Figures 3–6 illustrate the comparative F1-score performance of linear and nonlinear classifiers on Dataset 1 and Dataset 2, covering baseline models as well as entropy-based and KL divergence-based ensemble approaches, with subfigures distinguishing between linear and nonlinear settings.

Table 15. KL-inverse weighted ensemble weights of linear classifiers across embeddings for Dataset 1 (News Corpus) and Dataset 2 (CausalLM–Adjective group).

Classifier	Dataset 1						Dataset 2					
	TF-IDF	BoW	Word2Vec-CBOW	FastText	Skip-Gram	GloVe	TF-IDF	BoW	Word2Vec-CBOW	FastText	Skip-Gram	GloVe
Logistic Regression	0.42	0.39	0.36	0.37	0.36	0.38	0.41	0.39	0.37	0.38	0.37	0.38
Linear SVM	0.35	0.34	0.33	0.34	0.34	0.33	0.35	0.34	0.33	0.34	0.34	0.33
Naive Bayes	0.23	0.27	0.31	0.29	0.30	0.29	0.24	0.27	0.30	0.28	0.29	0.29

Table 16. KL-inverse weighted ensemble weights of nonlinear classifiers across embeddings for Dataset 1 (News Corpus) and Dataset 2 (CausalLM–Adjective group).

Classifier	Dataset 1						Dataset 2					
	TF-IDF	BoW	Word2Vec-CBOW	FastText	Skip-Gram	GloVe	TF-IDF	BoW	Word2Vec-CBOW	FastText	Skip-Gram	GloVe
SVM-RBF	0.40	0.38	0.45	0.42	0.43	0.41	0.37	0.39	0.36	0.35	0.37	0.36
Random Forest	0.27	0.29	0.25	0.26	0.25	0.27	0.28	0.26	0.27	0.28	0.27	0.28
Extra Trees	0.20	0.19	0.18	0.19	0.19	0.20	0.21	0.20	0.21	0.21	0.20	0.21
XGBoost	0.13	0.14	0.12	0.13	0.13	0.12	0.14	0.15	0.16	0.16	0.16	0.15

Table 17. Uncertainty-aware ensemble performance across different embeddings (Logistic Regression, Linear SVM, and Naive Bayes) using inverse-KL weighting on Dataset 1 and Dataset 2.

Embeddings	Dataset 1						Dataset 2					
	Accuracy	Precision	Recall	F1-Score	LogLoss	Mean KL	Accuracy	Precision	Recall	F1-Score	LogLoss	Mean KL
TF-IDF	0.8978	0.8988	0.8945	0.8962	0.4298	0.0966	0.845	0.848	0.841	0.844	0.392	0.210
BoW	0.8820	0.8840	0.8780	0.8800	0.4450	0.1100	0.835	0.837	0.832	0.834	0.400	0.220
Glove	0.8350	0.8370	0.8300	0.8330	0.4600	0.1400	0.820	0.823	0.818	0.820	0.415	0.240
SkipGram	0.8400	0.8420	0.8350	0.8380	0.4550	0.1350	0.823	0.825	0.820	0.822	0.412	0.238
Word2Vec-CBOW	0.8380	0.8400	0.8330	0.8350	0.4570	0.1370	0.821	0.823	0.819	0.821	0.414	0.239
FastText	0.8420	0.8440	0.8370	0.8400	0.4520	0.1320	0.825	0.827	0.822	0.824	0.410	0.236

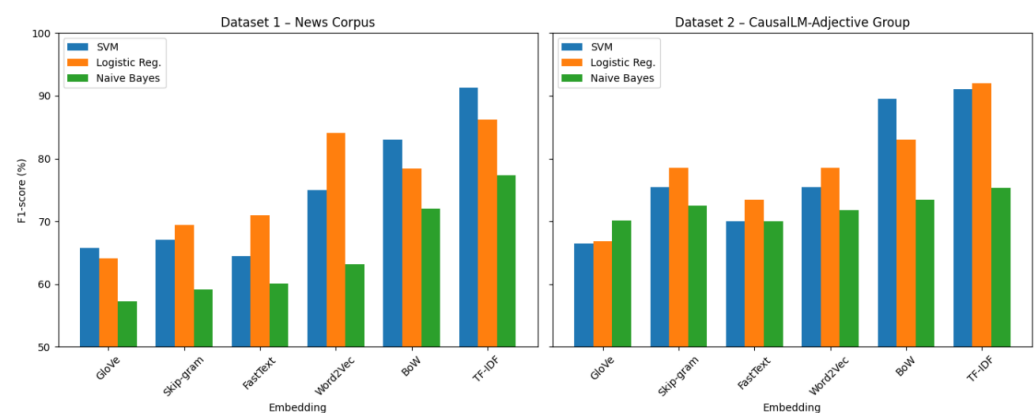
Note: Bold values indicate the best-performing embedding under the inverse-KL-weighted ensemble for each evaluation metric within the corresponding dataset.

Table 18. Uncertainty-aware ensemble performance across different embeddings on nonlinear classifiers (SVM-RBF, Random Forest, Extra Trees, and XGBoost) using inverse-KL weighting on Dataset 1 and Dataset 2.

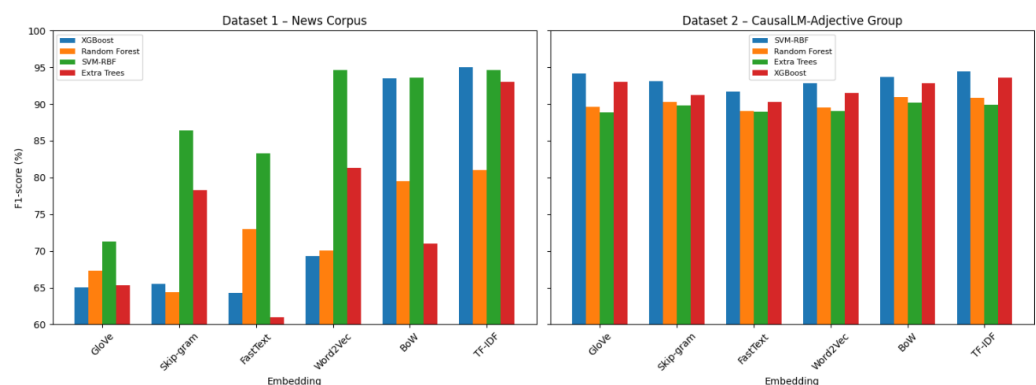
Embedding	Dataset 1						Dataset 2					
	Accuracy	Precision	Recall	F1-Score	LogLoss	KL Mean	Acc.	Prec.	Recall	F1	LogLoss	KL Mean
TF-IDF	0.961	0.959	0.952	0.956	0.292	0.072	0.976	0.974	0.970	0.972	0.221	0.055
BoW	0.945	0.943	0.936	0.940	0.318	0.089	0.969	0.967	0.962	0.965	0.236	0.061
Word2Vec-CBOW	0.927	0.925	0.918	0.921	0.339	0.104	0.961	0.959	0.954	0.956	0.247	0.067
FastText	0.933	0.931	0.924	0.928	0.331	0.096	0.964	0.962	0.957	0.959	0.242	0.063
Skip-gram	0.925	0.923	0.916	0.919	0.342	0.108	0.958	0.956	0.950	0.953	0.253	0.070
GloVe	0.913	0.911	0.904	0.907	0.356	0.121	0.951	0.949	0.943	0.946	0.264	0.076

Note: Bold values indicate the best-performing embedding under the inverse-KL-weighted ensemble for each evaluation metric within the corresponding dataset.

F1-score of Linear Classifiers on Dataset 1 and Dataset 2

**Figure 3.** F1-score of linear classifiers on Dataset 1 and Dataset 2.

F1-score of Non-Linear Classifiers on Dataset 1 and Dataset 2

**Figure 4.** F1-score of nonlinear classifiers on Dataset 1 and Dataset 2.

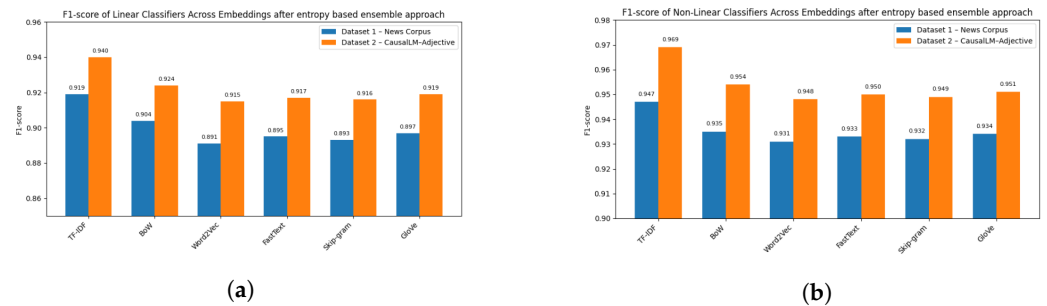


Figure 5. F1-score comparison on Dataset 1 and Dataset 2 of entropy-based ensemble classifiers. (a) Linear classifiers, (b) nonlinear classifiers.

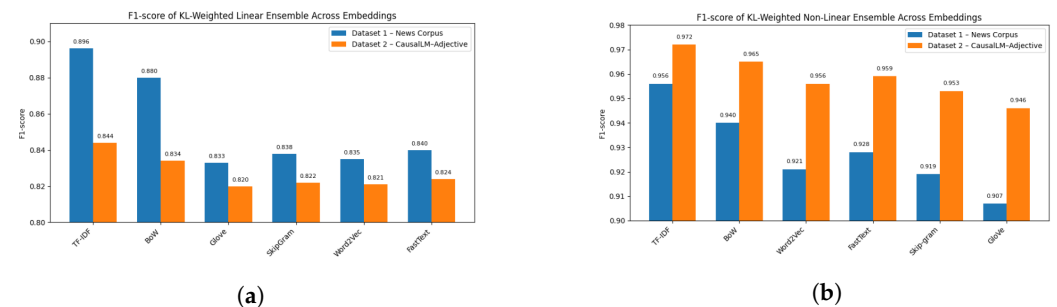


Figure 6. F1-score comparison on Dataset 1 and Dataset 2 of KL divergence-based ensemble classifiers. (a) Linear classifiers, (b) nonlinear classifiers.

5. Discussion

In this section, a detailed analysis of both entropy- and KL divergence-based ensemble approaches is presented for both linear and nonlinear classifiers. In the case of the linear classifier, the entropy-based ensemble proved to have more confident predictions with low entropy values as shown in Table 11. The linear classifiers generate a well-calibrated output probability with the simple and convex decision boundaries, and the entropy approach measures the spread of predictive probability vectors. As for linear models, they produce well-calibrated class probabilities, making it easier for optimizations, whereas for nonlinear classifiers, there is a high chance of generating overconfident and low calibrated probability scores, even with high accuracy noticed. So in this case, the KL divergence measure proved to be more effective than the entropy-based weighting. The prediction probability with the true label distribution, which is approximated by validation, is highly suitable for nonlinear classifiers, as it directly penalizes the mismatch by reducing the KL Mean and LogLoss. The confusion metrics provide insights into the discriminative ability and error distribution, which is critical for an uncertainty-aware system. The confusion metrics of each linear and nonlinear classifier showing how the prediction is accurate are shown in Figures 7–10.

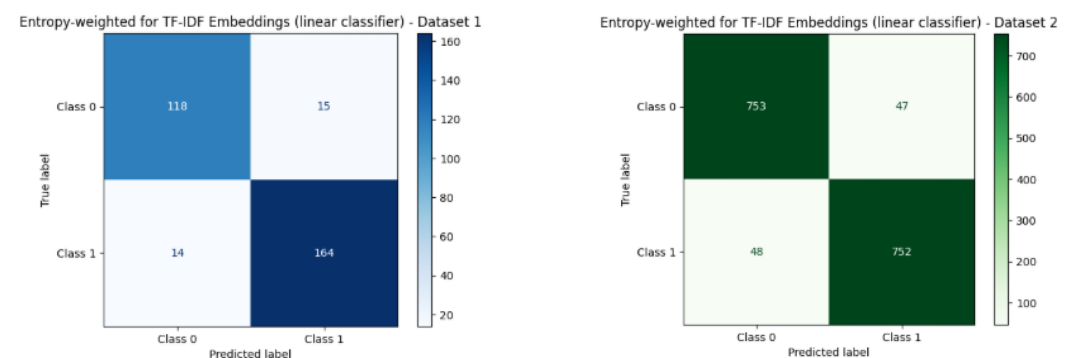


Figure 7. Entropy-weighted for TF-IDF embeddings (linear classifier)—Dataset 1 and Dataset 2.

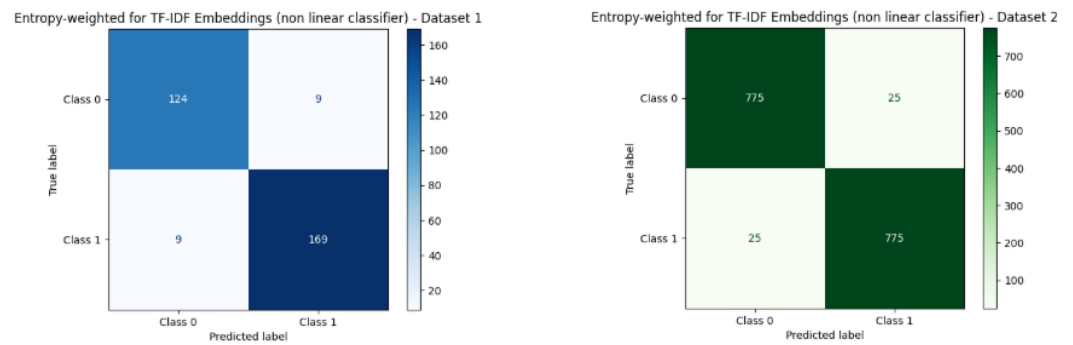


Figure 8. Entropy-weighted for TF-IDF embeddings (nonlinear classifier)—Dataset 1 and Dataset 2.

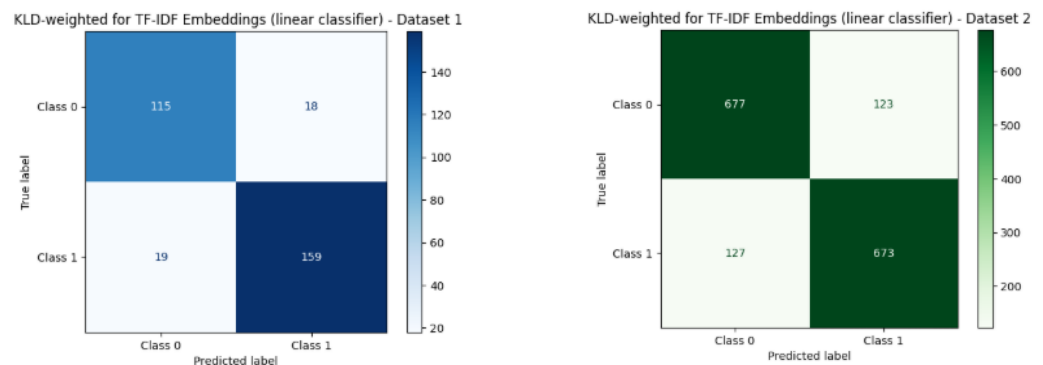


Figure 9. KLD-weighted for TF-IDF embeddings (linear classifier)—Dataset 1 and Dataset 2.

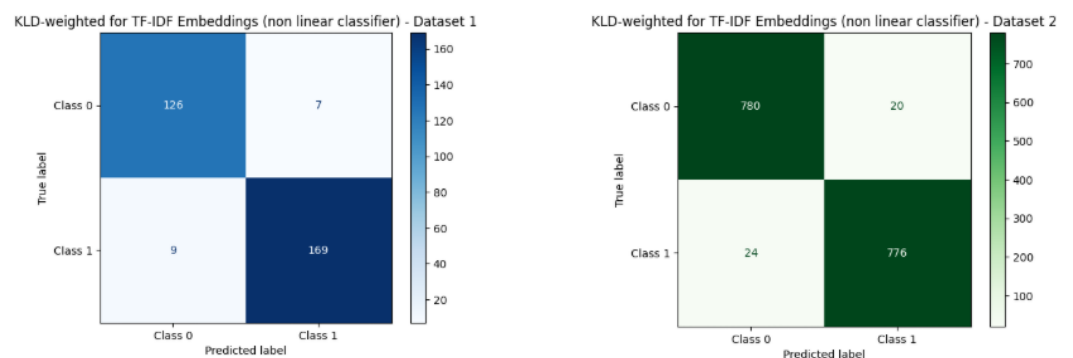


Figure 10. KLD-weighted for TF-IDF embeddings (nonlinear classifier)—Dataset 1 and Dataset 2.

The certain and uncertain predictions of the model's performance on the best classifiers in both the linear and nonlinear models is shown in the plots of Figures 11–14. The certain and uncertain predictions before and after the weighted ensemble approach visually highlight how the proposed approach improves the evaluation scores compared to the single classifiers. It is noticed to have a significant uplift in the result before and after the ensemble. This also showcases how the weighted ensemble approach provides a consistent and measurable improvement in both datasets on the linear and nonlinear approaches.

The entropy heatmap shown in Figure 15 illustrates the variability of nonlinear classifiers across different embeddings. A clear distinction can be seen across embeddings, with models trained on GloVe and Word2Vec-Skip-gram generally exhibiting higher entropy in Dataset 1, while BoW and TF-IDF yield comparatively lower values, indicating more stable prediction patterns. Random Forest and Extra Trees generally exhibit higher entropy with GloVe and Skip-gram embeddings, indicating lower confidence in prediction. On the other hand, TF-IDF consistently shows lower entropy values, particularly with Extra Trees, indicating more stable outcomes. These results highlight the sensitivity of nonlinear classifiers to the choice of embedding, and further justify the necessity for statistical ranking tests (e.g., Friedman and Nemenyi) to determine the relative significance of embeddings in causal text classification.

The statistical analysis further substantiates the observations from the entropy heatmaps. Table 19 displays the Friedman test statistics examining if the embedding choice significantly affects classifier stability. In case of Dataset 1, the test produced a Friedman statistic of 18.31 with a p -value of 0.0026, signifying a statistically relevant difference across embeddings at the 0.01 level. Table 20 presents the Nemenyi post hoc comparisons for Dataset 1. Nearly all embedding pairs differ significantly ($p < 0.05$), with BoW and TF-IDF showing the strongest and most consistent improvements. Both achieve significantly lower entropy than GloVe, Skip-gram, FastText, and Word2Vec-CBOW, confirming that BoW and TF-IDF provide the most stable representations.

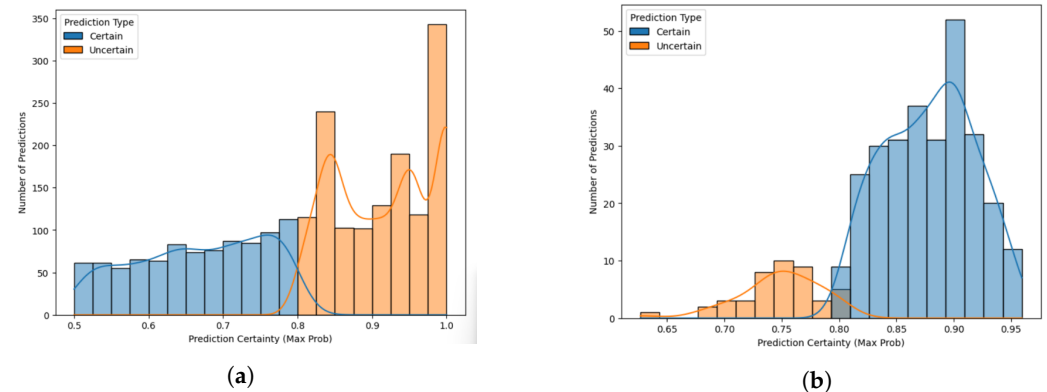


Figure 11. Certain and uncertain prediction possibilities using entropy weighting (linear classifier)—Dataset 1. (a) Before ensemble, (b) after ensemble.

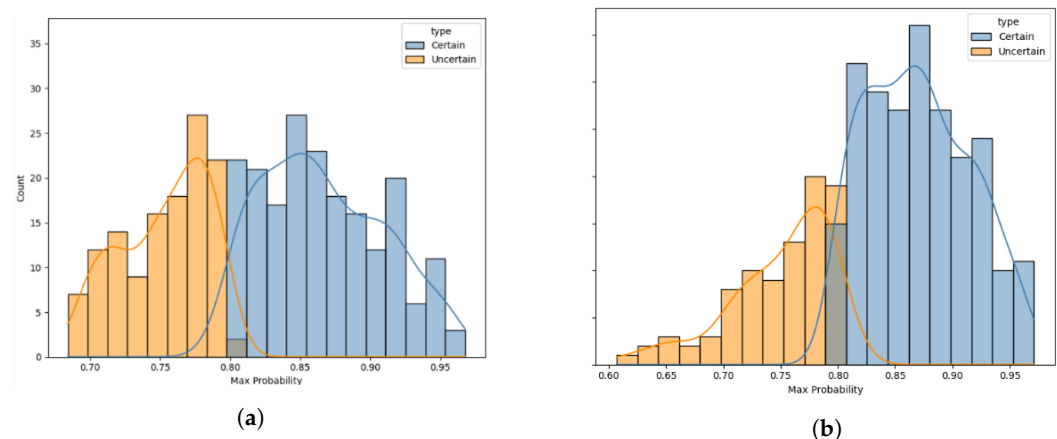


Figure 12. Certain and uncertain prediction possibilities using entropy weighting (linear classifier)—Dataset 2. (a) Before ensemble, (b) after ensemble.

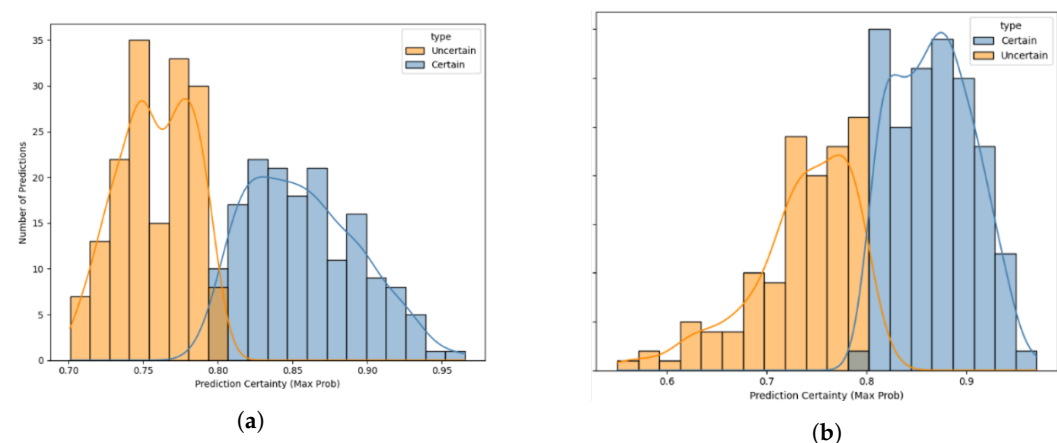


Figure 13. Certain and Uncertain prediction possibilities using KLD weighting (nonlinear classifier)—Dataset 1. (a) Before ensemble, (b) after ensemble.

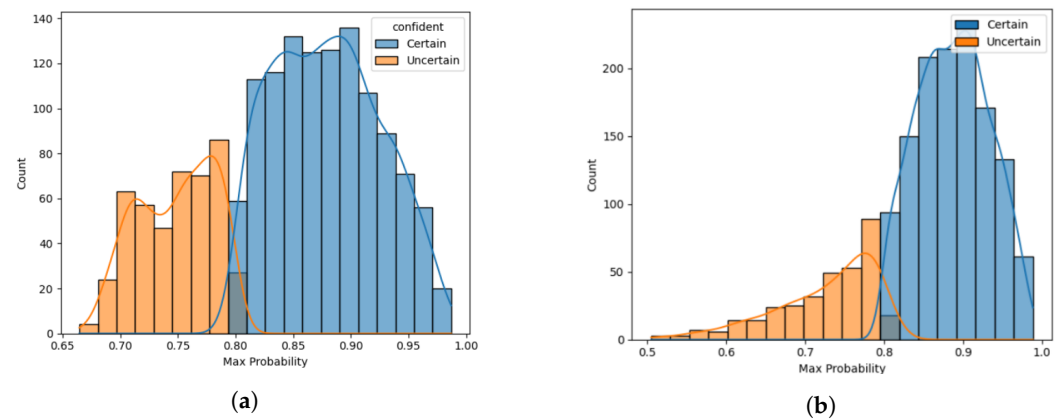


Figure 14. Certain and uncertain prediction possibilities using KLD weighting (nonlinear classifier)—Dataset 2. (a) Before ensemble, and (b) after ensemble.

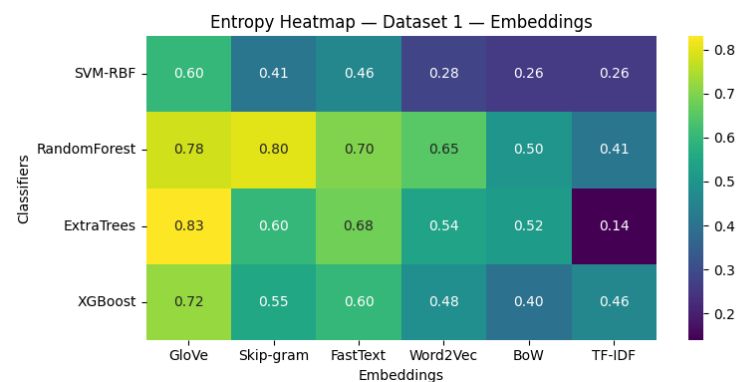


Figure 15. Entropy heatmaps of nonlinear classifiers (SVM-RBF, Random Forest, Extra Trees, and XGBoost) across different embeddings for Dataset 1.

Table 19. Friedman test statistics for embeddings across Dataset 1.

Dataset	Friedman Statistic	<i>p</i> -Value
Dataset 1	18.3094	0.0026

Table 20. Nemenyi post hoc pairwise comparisons of embeddings for Dataset 1. Shown are *p*-values; “Yes” indicates a significant difference at $\alpha = 0.05$.

	GloVe	Word2Vec-Skip-Gram	FastText	Word2Vec-CBoW	BoW	TF-IDF
TF-IDF	0.0387 (Yes)	0.0106 (Yes)	0.0257 (Yes)	0.0294 (Yes)	0.0032 (Yes)	1.0000 (No)
Word2Vec-CBoW	0.0098 (Yes)	0.0156 (Yes)	0.0262 (Yes)	1.0000 (No)	0.0217 (Yes)	0.0150 (Yes)
GloVe	1.0000 (No)	0.0190 (Yes)	0.0466 (Yes)	0.0361 (Yes)	0.0297 (Yes)	0.0085 (Yes)
Word2Vec-Skip-gram	0.0085 (Yes)	1.0000 (No)	0.0038 (Yes)	0.0426 (Yes)	0.0299 (Yes)	0.0350 (Yes)
BoW	0.0304 (Yes)	0.0077 (Yes)	0.0150 (Yes)	0.0186 (Yes)	1.0000 (No)	0.0032 (Yes)
FastText	0.0020 (Yes)	0.0476 (Yes)	1.0000 (No)	0.0410 (Yes)	0.0112 (Yes)	0.0197 (Yes)

Table 21 compares selected methods on Dataset 1. While baseline approaches achieved F1-scores between 83 and 88%, the proposed uncertainty-weighted ensemble substantially outperformed all, reaching 95.6%, highlighting its robustness and effectiveness in causal relation detection.

Table 21. Comparative analysis of the proposed weighted ensemble models—entropy-based (linear) and KL divergence-based (nonlinear)—with existing methods on Dataset 1.

Author(s)	Model Used	Results
Tan et al. (2022) [13]	BERT-Base-cased	F1: 83.46% (5-fold CV)
Patel (2023) [14]	SuperCon	F1: 89.72%
Schrader et al. (2023) [15]	RoBERTa	F1 (binary & span): 88.30%
Burdisso et al. (2022) [16]	Ensemble-10m	F1: 85.08%
Adibhatla et al. (2022) [17]	RoBERTa-base+Adapter	F1 (Subtask 1): 85.30%, F1 (Subtask 2): 83.20%
Proposed Entropy-based Weighted Ensemble (Linear)	TF-IDF - Ensemble Linear	F1: 91.9% (10-fold CV)
Proposed KL divergence-based Weighted Ensemble (Nonlinear)	TF-IDF - Ensemble Nonlinear	F1: 95.6% (10-fold CV)

Note: Bold rows highlight the proposed ensemble models and their corresponding best-performing results on Dataset 1.

Tables 22 and 23 show some examples of TF-IDF-based linear classifiers (Linear SVM, Logistic Regression, and Naive Bayes) on text instances where individual models misclassified the input. Table 22 highlights cases where the entropy-based weighted ensemble was able to correct the errors of one or more base classifiers, showing its ability to correct the misclassified instances. In contrast, Table 23 lists cases where the weighted ensemble model also failed, indicating limitations in handling certain ambiguous and challenging examples. Similarly, Tables 24 and 25 show the performance of nonlinear classifiers (SVM-RBF, Random Forest, Extra Trees, and XGBoost) with a KL divergence-based ensemble. While the ensemble successfully resolved some misclassifications in Table 24, Table 25 illustrates instances where even the ensemble could not get a correct prediction, highlighting persistent challenges in text classification tasks despite uncertainty-aware methodologies. The code of all these experiments is available at <https://github.com/sivachandrakb/Uncertainty-Weighted-Ensemble-Learning> (accessed on 2 February 2026).

Table 22. Failed examples of TF-IDF-based linear classifiers (Linear SVM, Logistic Regression, and Naive Bayes) where individual models misclassified the instance but were corrected by the entropy-based weighted ensemble. Probabilities are given as [P(class 0), P(class 1)].

Text (Truncated)	True Label	LinearSVM Pred	LinearSVM [p0, p1]	LogisticRegression Pred	LogisticRegression [p0, p1]	NaiveBayes Pred	NaiveBayes [p0, p1]	Ensemble Pred	Ensemble [p0, p1]	Failed Models
the march descended into chaos...	1	0	[0.509, 0.491]	0	[0.522, 0.478]	1	[0.576, 0.424]	1	[0.514, 0.486]	LinearSVM, LogisticRegression
two militants had also been killed in the fire fight	1	0	[0.502, 0.498]	1	[0.530, 0.470]	1	[0.645, 0.355]	1	[0.556, 0.444]	LinearSVM
the alliance claimed took part last year	0	0	[0.830, 0.170]	0	[0.722, 0.278]	1	[0.516, 0.484]	0	[0.683, 0.317]	NaiveBayes
they said the suspects had initially planned to kill...	0	0	[0.639, 0.361]	1	[0.504, 0.496]	1	[0.553, 0.447]	0	[0.530, 0.470]	LogisticRegression, NaiveBayes

Table 23. Failed examples of TF-IDF-based linear classifiers (Linear SVM, Logistic Regression, and Naive Bayes) where individual models misclassified the instance and entropy-based weighted ensemble fails. Probabilities are given as [P(class 0), P(class 1)].

Text (Truncated)	True Label	LinearSVM Pred	LinearSVM [p0, p1]	LogisticRegression Pred	LogisticRegression [p0, p1]	NaiveBayes Pred	NaiveBayes [p0, p1]	Ensemble Pred	Ensemble [p0, p1]	Failed Models
it is somehow like a revolution...	1	0	[0.511, 0.489]	1	[0.538, 0.462]	0	[0.530, 0.470]	0	[0.501, 0.499]	
testifying before the farlam commission of inquiry...	0	1	[0.509, 0.491]	1	[0.515, 0.485]	1	[0.525, 0.475]	1	[0.516, 0.484]	
nzuza earlier testified he was initially part of the national union of mineworkers...	1	0	[0.777, 0.223]	0	[0.687, 0.313]	0	[0.545, 0.455]	0	[0.673, 0.327]	
thirty-four people, mostly striking mineworkers, were shot dead in a clash with police	1	0	[0.659, 0.341]	0	[0.591, 0.409]	1	[0.541, 0.459]	0	[0.573, 0.427]	

Table 24. Examples of TF-IDF-based nonlinear classifiers (SVM-RBF, Random Forest, Extra Trees, and XGBoost) where the KL divergence-based ensemble model passed. Probabilities are given as [P(class 0), P(class 1)].

Text (Truncated)	True Label	SVM-RBF Pred	SVM-RBF [p0, p1]	RF Pred	RF [p0, p1]	ExtraT Pred	ExtraT [p0, p1]	XGB Pred	XGB [p0, p1]	Ensemble Pred	Ensemble [p0, p1]	Failed Models
in the other gun battle...	1	0	[0.602, 0.398]	0	[0.530, 0.470]	0	[0.615, 0.385]	1	[0.172, 0.828]	1	[0.478, 0.522]	SVM-RBF, RF, ExtraT
psu insurance employees protest...	1	0	[0.561, 0.439]	1	[0.180, 0.820]	1	[0.205, 0.795]	1	[0.068, 0.932]	1	[0.254, 0.746]	SVM-RBF
they joined the protesters...	0	0	[0.500, 0.500]	1	[0.465, 0.535]	1	[0.465, 0.535]	0	[0.848, 0.152]	0	[0.572, 0.428]	RF, ExtraT
it's already been two months...	1	0	[0.529, 0.471]	1	[0.445, 0.555]	1	[0.385, 0.615]	1	[0.359, 0.641]	1	[0.430, 0.570]	SVM-RBF

Table 25. Examples of TF-IDF-based nonlinear classifiers (SVM-RBF, Random Forest, Extra Trees, and XGBoost) where the KL divergence-based ensemble model also failed. Probabilities are given as [P(class 0), P(class 1)].

Text (Truncated)	True Label	SVM-RBF Pred	SVM-RBF [p0, p1]	RF Pred	RF [p0, p1]	ExtraT Pred	ExtraT [p0, p1]	XGB Pred	XGB [p0, p1]	Ensemble Pred	Ensemble [p0, p1]	Failed Models
tamil nadu activists support medha patkar...	1	1	[0.437, 0.563]	0	[0.510, 0.490]	1	[0.380, 0.620]	0	[0.712, 0.288]	0	[0.511, 0.489]	
it is learnt that many leaders of auto rickshaw trade unions...	1	0	[0.725, 0.275]	1	[0.275, 0.725]	1	[0.375, 0.625]	0	[0.590, 0.410]	1	[0.494, 0.506]	
tuen mun court deputy magistrate so kaiche-ung...	1	0	[0.570, 0.430]	1	[0.465, 0.535]	1	[0.415, 0.585]	0	[0.752, 0.248]	0	[0.552, 0.448]	
however, he called off the fast after getting a call...	1	1	[0.443, 0.557]	0	[0.555, 0.445]	1	[0.490, 0.510]	0	[0.625, 0.375]	0	[0.529, 0.471]	

6. Conclusions and Future Scope

In this work, we propose an uncertainty-weighted ensemble system for the detection of causal text, which combines various embeddings and classifiers with dynamic weighting mechanisms based on entropy and KL divergence. This method gives importance to the models that show confident predictions while decreasing the importance of uncertain ones. Experiments were carried out on two benchmarked datasets. Results show that the ensemble-aware weighting makes the evaluation scores better. It can be seen that the entropy-based weighting is better for linear classifiers, while KL divergence weighting is better for nonlinear classifiers, achieving higher LogLoss and calibration scores. Adding uncertainty estimation to ensemble decision-making helps to reduce overconfidence and improve trustworthy predictions. While the approach shows high predictive performance focusing on traditional ML models, future work could look into uncertainty-aware ensemble learning using DL and transformers, and this would be a strong next step in future work.

Author Contributions: S.K.B. contributed to the investigation, software implementation, and writing—original draft preparation. N.M. contributed to the methodology, validation, and investigation. M.K.K. contributed to validation, investigation, and writing—review and editing. S.O.K. contributed to conceptualization, validation, and writing—review and editing. S.K.S. contributed to problem conceptualization, methodology development, supervision of the research work, and writing—review and editing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was not funded by any external agency.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available. The Causal News Corpus can be accessed at <https://github.com/tanfiona/CausalNewsCorpus/tree/master/data/V1> (accessed on 2 February 2026), and the CausalM dataset is available at <https://www.kaggle.com/datasets/amirfeder/causalm> (accessed on 2 February 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hu, M.; Zhang, Z.; Zhao, S.; Huang, M.; Wu, B. Uncertainty in Natural Language Processing: Sources, Quantification, and Applications. *arXiv* **2023**, arXiv:2306.04459. [CrossRef]
2. Baan, J.; Daheim, N.; Ilia, E.; Ulmer, D.; Li, H.-S.; Fernández, R.; Plank, B.; Sennrich, R.; Zerva, C.; Aziz, W. Uncertainty in Natural Language Generation: From Theory to Applications. *arXiv* **2023**, arXiv:2307.15703. [CrossRef]
3. Xiao, Y.; Wang, W.Y. Quantifying Uncertainties in Natural Language Processing Tasks. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 7322–7329. [CrossRef]
4. Ulmer, D.; Frellsen, J.; Hardmeier, C. Exploring Predictive Uncertainty and Calibration in NLP: A Study on the Impact of Method & Data Scarcity. In *Findings of the Association for Computational Linguistics: EMNLP 2022*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022.
5. Hu, Y.; Khan, L. Uncertainty-aware reliable text classification. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Virtual, 14–18 August 2021; pp. 628–636.
6. Prasad, R.; Dinesh, N.; Lee, A.; Mitsakaki, E.; Robaldo, L.; Joshi, A.; Webber, B. The Penn Discourse TreeBank 2.0. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco, 26 May–1 June 2008.
7. Miwa, M.; Thompson, P.; Ananiadou, S. BioCause: Annotating and analysing causal relations in the biomedical domain. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 23–25 May 2012.
8. Hendrickx, I.; Kim, S.N.; Kozareva, Z.; Semmar, N.; Szpakowicz, S.; Zanzotto, F.M. SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals. In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval), Los Angeles, CA, USA, 15–16 July 2010.

9. Gao, T.; Meng, Y.; Chen, D. CAUSALBERT: Injecting causal knowledge into pretrained language models for causal relation extraction. In Proceedings of the Findings of the Association for Computational Linguistics (ACL), Virtual, 1–6 August 2021.
10. Girju, R. Automatic detection of causal relations for question answering. In Proceedings of the ACL Workshop on Multilingual Summarization and Question Answering, Sapporo, Japan, 11 July 2003.
11. Straka, M.; Straková, J. UDPipe: Trainable pipeline for processing CoNLL-U files. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Portorož, Slovenia, 23–28 May 2016.
12. Straka, M. UDPipe 2.0: Models and design. In *CoNLL Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018.
13. Tan, F.A.; Hürriyetoglu, A.; Caselli, T.; Oostdijk, N.; Nomoto, T.; Hettiarachchi, H.; Ameer, I.; Uca, O.; Liza, F.F.; Hu, T. The causal news corpus: Annotating causal relations in event sentences from news. *arXiv* **2022**, arXiv:2204.11714. [\[CrossRef\]](#)
14. Patel, R. Interosml@ causal news corpus 2023: Understanding causal relationships: Supervised contrastive learning for event classification. In Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, Varna, Bulgaria, 7 September 2023; pp. 60–65.
15. Schrader, T.P.; Razniewski, S.; Lange, L.; Friedrich, A. Boschai@ causal news corpus 2023: Robust cause-effect span extraction using multi-layer sequence tagging and data augmentation. *arXiv* **2023**, arXiv:2312.06338.
16. Burdisso, S.; Zuluaga-Gomez, J.; Villatoro-Tello, E.; Fajcik, M.; Singh, M.; Smrz, P.; Motlicek, P. IDIAPers@ causal news corpus 2022: Efficient causal relation identification through a prompt-based few-shot approach. *arXiv* **2022**, arXiv:2209.03895.
17. Adibhatla, H.S.; Shrivastava, M. LTRC@ causal news corpus 2022: Extracting and identifying causal elements using adapters. In Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text (CASE), Abu Dhabi, United Arab Emirates, 7–8 December 2022; pp. 50–55.
18. Jia, L.; Yang, F.; Man, K.L.; Purwanto, E.; Guan, S.-U.; Smith, J.; Yue, Y. Quantify the Causes of Causal Emergence: Critical Conditions of Uncertainty and Asymmetry in Causal Structure. *arXiv* **2024**, arXiv:2212.01551. [\[CrossRef\]](#)
19. Chau, S.L.; Ton, J.-F.; González, J.; Teh, Y.W.; Sejdinovic, D. BAYESIMP: Uncertainty Quantification for Causal Data Fusion. Unpublished Manuscript/Preprint, University of Oxford and Collaborators. *arXiv* **2021**, arXiv:2106.03477.
20. Annadani, Y.; Rothfuss, J.; Lacoste, A.; Scherrer, N.; Goyal, A.; Bengio, Y.; Bauer, S. Variational Causal Networks: Approximate Bayesian Inference over Causal Structures. *arXiv* **2021**, arXiv:2106.07635. [\[CrossRef\]](#)
21. Morucci, M.; Noor-E-Alam, M.; Rudin, C. *A Robust Approach to Quantifying Uncertainty in Matching Problems of Causal Inference*; Technical Report; Duke University: Durham, NC, USA; Northeastern University: Boston, MA, USA, 2022.
22. Chaudhry, A.; Xu, P.; Gu, Q. Uncertainty Assessment and False Discovery Rate Control in High-Dimensional Granger Causal Inference. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.
23. Jesson, A.; Mindermann, S.; Shalit, U.; Gal, Y. Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models. *arXiv* **2020**, arXiv:2007.00163.
24. Atmaja, B.T.; Sasou, A.; Burkhardt, F. Uncertainty-Based Ensemble Learning for Speech Classification. In *Proceedings of the 27th Conference of the Oriental COCOSDA (O-COCOSDA)*, Hsinchu City, Taiwan, 17–19 October 2024; IEEE: New York, NY, USA, 2024; pp. 1–6.
25. Ruan, Y.; Singh, S.; Morningstar, W.; Alemi, A.A.; Ioffe, S.; Fischer, I.; Dillon, J.V. Weighted Ensemble Self-Supervised Learning. *arXiv* **2022**, arXiv:2211.09981.
26. Aravind Sekhar, R.; Sreeni, K.G. An Entropy Based Alternate Approach to Develop an Ensemble Classifier. *Indian J. Sci. Technol.* **2023**, *16*, 2601–2608. [\[CrossRef\]](#)
27. Wang, J.; Xu, S.; Duan, B.; Liu, C.; Liang, J. An Ensemble Classification Algorithm Based on Information Entropy for Data Streams. *arXiv* **2017**, arXiv:1708.03496. [\[CrossRef\]](#)
28. Al-Maadeed, R.A.-M.; Hasaneen, M.H. Entropy-Based Ensemble of Convolutional Neural Networks for Clothes Texture Pattern Recognition. *Appl. Sci.* **2022**, *14*, 10730.
29. Malinin, A.; Gales, M. Reverse KL-Divergence Training of Prior Networks: Improved Uncertainty and Adversarial Robustness. In *Advances in Neural Information Processing Systems*; NeurIPS: San Diego, CA, USA, 2019; Volume 32.
30. Sensoy, M.; Kaplan, L.; Kandemir, M. Evidential Deep Learning to Quantify Classification Uncertainty. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 3–8 December 2018.
31. Kumar, N.; Narang, A.; Lall, B. Kullback–Leibler Divergence-Based Regularized Normalization for Low-Resource Tasks. *IEEE Trans. Artif. Intell.* **2023**, *5*, 2638–2650. [\[CrossRef\]](#)
32. Lee, K.; Lee, H.; Lee, K.; Shin, J. Training Confidence-Calibrated Classifiers for Detecting Out-of-Distribution Samples. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
33. Mohammed, A.; Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 757–774. [\[CrossRef\]](#)

34. Azim, K.; Tahir, A.; Shahroz, M.; Karamti, H.; Almeyda Vazquez, A.; Rojas Vistorte, A.; Ashraf, I. Ensemble stacked model for enhanced identification of sentiments from IMDB reviews. *Sci. Rep.* **2025**, *15*, 13405. [[CrossRef](#)]
35. Kavitha, M.; Prabhavathy, P. A hybrid approach for text classification. *Int. J. Res. Anal. Rev.* **2023**, *10*, 234–238.
36. Ashiga, M.; Jie, W.; Wu, F.; Voskanyan, V.; Dinmohammadi, F.; Brookes, P.; Gong, J.; Wang, Z. Ensemble learning for large language models in text and code generation. *arXiv* **2025**, arXiv:2503.13505.
37. Ibrahim, Y.; Okafor, E.; Yahaya, B.; Yusuf, S.M.; Abubakar, Z.M.; Bagaye, U.Y. Comparative study of ensemble learning techniques for text classification. In Proceedings of the 2021 1st International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS), Abuja, Nigeria, 15–16 July 2021.
38. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [[CrossRef](#)]
39. Harris, Z. Distributional structure. *Word* **1954**, *10*, 146–162. [[CrossRef](#)]
40. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. In Proceedings of the International Conference on Learning Representations (ICLR), Scottsdale, AZ, USA, 2–4 May 2013.
41. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [[CrossRef](#)]
42. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
43. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
44. Hosmer, D.W.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*, 3rd ed.; John Wiley & Sons: New York, NY, USA, 2013.
45. McCallum, A.; Nigam, K. A comparison of event models for Naive Bayes text classification. In Proceedings of the AAAI-98 Workshop on Learning for Text Categorization, Bethesda, MD, USA, 2–7 November 1998.
46. Schölkopf, B.; Smola, A.J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2002.
47. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
48. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [[CrossRef](#)]
49. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
50. Tan, F. CausalNewsCorpus. GitHub Repository. 2023. Available online: <https://github.com/tanfiona/CausalNewsCorpus/tree/master/data/V1> (accessed on 2 February 2026).
51. Feder, A.; Oved, N.; Shalit, U.; Reichart, R. CausaLM: Causal Model Explanation through Counterfactual Language Models. *Comput. Linguist.* **2021**, *47*, 333–386. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.