Exploring Causal Directions through Word Occurrences: Semi-supervised Bayesian Classification Framework

King Tao Jason Ng and **Diego Mollá**
School of Computing
Macquarie University
Sydney, Australia
{kingtao.ng,diego.molla-aliod}@mq.edu.au

Abstract

Determining causal directions in sentences plays a critical role into understanding a cause-and-effect relationship between entities. In this paper, we show empirically that word occurrences from several Internet domains resemble the characteristics of causal directions. Our research contributes to the knowledge of the underlying data generation process behind causal directions. We propose a two-phase method: 1. Bayesian framework, which generates synthetic data from posteriors by incorporating word occurrences from the Internet domains. 2. Pre-trained BERT, which utilises semantics of words based on the context to perform classification. The proposed method achieves an improvement in performance for the Cause-Effect relations of the SemEval-2010 dataset, when compared with random guessing.

1 Introduction

Understanding causality is critical for various tasks including Question Answering. Singer et al. (1992) provide a great example: *Dorothy poured water on the fire. The fire went out*. Subsequently, if it is followed by the question *did she put out the fire?*, the answer is *yes* because *poured water on* implies that the two sentences are causally linked.

When provided with two entities, namely $e1$ and $e2$, in the sentence that are known to have a causal relation, the causal direction tells us which one is a cause and which one is an effect. In the previous example, *poured water on* is the cause whereas *fire went out* is the effect. Therefore, the causal direction in this case is *poured water on* $\rightarrow$ *fire went out*.

In this study, we show that word occurrences resemble the characteristics of causal directions. Our research contributes to the knowledge of the underlying data generation process behind causal directions. To achieve this, we propose a semi-supervised classification method[1] for determining a causal direction if its causal relation is known to exist in the sentence. The GitHub page[2] is available for reference purposes.

2 Related Work

In this section, we provide a brief overview of two approaches for identifying causal relations. The first approach, Pointwise Mutual Information, is designed to eliminate the need for corpus creation. The second approach, Data Augmentation, clearly involves the need of creating a corpus.

Pointwise Mutual Information. If $e1$ and $e2$ are causally related, it is expected that they will frequently appear together (Kroeger, 2005). Pointwise Mutual Information (PMI) (Glickman et al., 2005) is a notable measure used to assess co-occurrence. However, it should be noted that PMI is commutative and therefore it cannot distinguish between the causal directions $e1 \rightarrow e2$ and $e2 \rightarrow e1$. Let us say two entities, $e1$ and $e2$. Suppes (1973) points out $e1$ is a possible cause of $e2$ if $e2$ is mentioned more frequently with $e1$ than by itself.

$$P(e2 \mid e1) > P(e2) \tag{1}$$

We rewrite Equation (1) as follows:

$$\frac{P(e2 \cap e1)}{P(e1)P(e2)} > 1 \tag{2}$$

Equation (2) is elegant if $e1$ and $e2$ establish a causal relation, but it fails to determine its causal direction. For example, if $e2$ is a cause of $e1$, we have

$$P(e1 \mid e2) > P(e1) \tag{3}$$

[1] Utilizing word occurrences to infer causal directions can be regarded as a form of supervised learning although it may be considered as a semi-supervised learning because labels are not annotated.

[2] https://github.com/kingtaojasonng/Causal_Direction

Proceedings of The 21st Annual Workshop of the Australasian Language Technology Association, pages 30–39
November 29 - December 1, 2023

After a couple of algebraic manipulations, we end up with

$$\frac{P(e1 \cap e2)}{P(e2)P(e1)} > 1 \quad (4)$$

Equations (2) and (4) are now identical. That is, we cannot distinguish $e1 \rightarrow e2$ from $e2 \rightarrow e1$ using PMI. This means that the same PMI equation is obtained regardless of the causal direction. Despite this limitation, PMI is commonly employed in the identification of causal relations (Moghimifar et al., 2020).

Data Augmentation. This is a prevalent strategy employed by many language models to address the difficulties posed by scenarios where there is a limited amount of labelled training data. To illustrate, Li et al. (2021) leverage external sources like CausalBank and ConceptNet to incorporate causal knowledge into pre-trained language models. It is worth noting that, even though they capture causal knowledge, there remains a need for human annotation in this process. The use of word occurrences, which is unannotated data, is a more cost-effective approach that can generalise to various scenarios.

3 Dataset

For our study, we use the SemEval-2010 (Task 8) dataset (Hendrickx et al., 2010). This dataset focuses on a multi-class classification task. However, for the purpose of our study, we narrow our attention to the specific category labelled as Cause-Effect in the dataset.

A sentence is considered as Cause-Effect if two entities, which are marked as <e1> and <e2>, show a causal relation.

```
"<e1>Suicide</e1> is one of the leading
    causes of <e2>death</e2> among pre-
    adolescents and teens, and victims
    of bullying are at an increased risk
     for committing suicide."
Cause-Effect(e1,e2)
```

Example 1: A sample sentence. The last line indicates *suicide* $\rightarrow$ *death*.

The Cause-Effect category comprises a total of 1,331 instances, divided between the training and test data. In the training data, there are 1,003 instances labelled as Cause-Effect, with 659 of them demonstrating the relationship $e2 \rightarrow e1$. In the test data, out of the 328 Cause-Effect instances, 134 exhibit the $e1 \rightarrow e2$ relationship. There are no bidirectional causal relations[3] in the dataset. Table 1 provides a summary of the SemEval-2010 (Task 8) dataset and Example 1 shows an example, which is taken from the training data.

Datasets		SemEval-2010 (Task 8) Raw Count	Percentage
Training	$e1 \rightarrow e2$	344	34.30%
	$e2 \rightarrow e1$	659	65.70%
	Total	1,003	100.00%
Test	$e1 \rightarrow e2$	134	40.85%
	$e2 \rightarrow e1$	194	59.15%
	Total	328	100.00%

Table 1: The distribution of SemEval-2010 (Task 8) is shown.

4 Method

In order to gain insights into the similarity between word occurrences and causal directions, we simulate a semi-supervised classification setup and *exclude* the training data from our analysis. The motivation behind examining word occurrences is that if two words frequently collocate, this linguistic clue can be used to infer a causal direction. For instance, if the words *smoking* and *lung cancer* frequently collocate, this pair suggests a potential causal direction, the direction of which we need to determine. Our method consists of two phases — Bayesian framework, and Pre-trained BERT.

4.1 Phase 1: Bayesian Framework

We propose a Bayesian framework that incorporates word occurrences from several Internet domains as priors. By leveraging the externally sourced data, this framework can generate synthetic data that exhibits similarities with causal directions.

Given two entities, namely $e1$ and $e2$, the direction of causality will be either $e1 \rightarrow e2$ or $e2 \rightarrow e1$. We formulate the problem definition into a hypothesis test by specifying the null and alternative hypotheses in the framework as shown in (5):

$$H_0 : \overbrace{f(e1 \rightarrow e2 \mid \mathbf{X})}^{\text{Model 1}} > \overbrace{f(e2 \rightarrow e1 \mid \mathbf{X})}^{\text{Model 2}} \quad (5)$$
$$H_a : \text{Otherwise}$$

where $\mathbf{X}$ is the training data, and f represents a

[3] An illustration of bidirectional causal relations is *the-chicken-or-the-egg* causal dilemma, which states chickens hatch from eggs and eggs are laid by chickens.

probability distribution[4]. The null hypothesis H_0 states that the density of $f(e1 \rightarrow e2 \mid \mathbf{X})$ mostly centres at an upper end of probability relative to $f(e2 \rightarrow e1 \mid \mathbf{X})$. Using the Bayes' rule, as shown in (6) and (7):

$$f(e1 \rightarrow e2 \mid \mathbf{X}) = \frac{f(\mathbf{X} \mid e1 \rightarrow e2) f(e1 \rightarrow e2)}{f(\mathbf{X})} \tag{6}$$

$$f(e2 \rightarrow e1 \mid \mathbf{X}) = \frac{f(\mathbf{X} \mid e2 \rightarrow e1) f(e2 \rightarrow e1)}{f(\mathbf{X})} \tag{7}$$

we re-write the null hypothesis as (8):

$$H_0 : f(\mathbf{X} \mid e1 \rightarrow e2) f(e1 \rightarrow e2) > f(\mathbf{X} \mid e2 \rightarrow e1) f(e2 \rightarrow e1) \tag{8}$$

Because no training data $\mathbf{X}$ is provided, we further simplify the null hypothesis as (9):

$$H_0 : f(e1 \rightarrow e2) > f(e2 \rightarrow e1) \tag{9}$$

This means that the posterior distributions are effectively the priors.

4.1.1 Priors

Since we exclude the training data, it is necessary to find a proxy for the causal direction. Broadly speaking, priors can be any type of information that conveys the knowledge of $f(e1 \rightarrow e2)$ and $f(e2 \rightarrow e1)$.

We use word occurrences from several Internet domains as priors to model causal directions. As SemEval-2010 (Task 8) is mainly extracted from Wikipedia, we select a wide range of the Internet domains, as shown in Table 2. These include media outlets, since Wikipedia often references news articles for the news; educational institutions which Wikipedia cites as learning resources; government entities, which Wikipedia references to gather information about agencies and policies; scientific publishers, which Wikipedia references for scientific knowledge; online resources that often link to Wikipedia pages for additional information; journals, where Wikipedia may reference the works of researchers and scholars; and general reference.

[4]A probability distribution is a mathematical function that describes the likelihoods of all possible outcomes that a random variable can take. Probability distributions not only allow us to quantify uncertainty but also provide a comprehensive view of all possible values and their associated probabilities. Hence, we employ Bayesian statistics as opposed to frequentist statistics in hypothesis testing to harness these advantages.

abc.net.au	au.news.yahoo.com	bbc.com
economist.com	edu	gov.au
imdb.com	mit.edu	nationalgeographic.com
ncbi.nlm.nih.gov	nejm.org	nytimes.com
oreilly.com	skynews.com.au	smh.com.au
springer.com	time.com	wikipedia.org
wiley.com		

Table 2: The Internet domains used for extracting word occurrences.

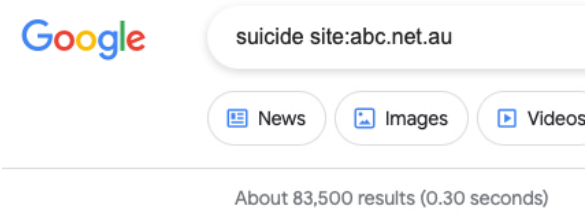


Figure 1: When searching for the word *suicide*, 83,500 results are shown.

We use Google search to determine word occurrences by restricting the search to these chosen Internet domains. For example, to look for the word *suicide* in the ABC News, the search command would be `suicide site:abc.net.au`, as shown in Figure 1. A number of search results (i.e., 83,500), which we consider occurrences, is shown before actual results are displayed. Figure 1 is for illustration purposes only. In practice, we searched Google programmatically.

To compute $P(e1 \rightarrow e2)$, which is a single probability, we use Google to estimate the frequency count of the occurrences of both $e1$ and $e2$ in a domain, $C(e1, e2)$, and divide it by the frequency count of the occurrences of $e1$ alone in the same domain, $C(e1)$. $P(e2 \rightarrow e1)$ is calculated using the same method. This will result in unnormalised versions, which will be normalised as described below.

$$P'(e1 \rightarrow e2) = \frac{C(e1, e2)}{C(e1)} \tag{10}$$

$$P'(e2 \rightarrow e1) = \frac{C(e1, e2)}{C(e2)} \tag{11}$$

In (10) and (11), $C(e1) \neq 0$ and $C(e2) \neq 0$ to avoid zero counts[5]. To normalise Equations (10)

[5]Haldane (1956) suggests adding 0.5 to every count if $C(e1) = 0$ or $C(e2) = 0$. However, we did not experience zero counts during the experiments.

and (11), both are divided by their sum.[6]

$$P(e1 \rightarrow e2) = \frac{P'(e1 \rightarrow e2)}{P'(e1 \rightarrow e2) + P'(e2 \rightarrow e1)} \quad (12)$$

$$P(e2 \rightarrow e1) = \frac{P'(e2 \rightarrow e1)}{P'(e1 \rightarrow e2) + P'(e2 \rightarrow e1)} \quad (13)$$

Equations (12) and (13) are effectively conditional probabilities. We apply Equations (12) and (13) repeatedly for each domain outlined in Table 2. This process results in two distinct lists of probabilities. Each of these lists provides a complete range of likelihoods. A probability[7] is assigned to each likelihood, effectively quantifying the uncertainty. As a result, we have $f(e1 \rightarrow e2)$ and $f(e2 \rightarrow e1)$.

Prior Specification

Prior specification is a process of selecting and defining a prior distribution in the Bayesian framework. More specifically, it involves choosing a type of distributions and its parameters to fit $f(e1 \rightarrow e2)$ and $f(e2 \rightarrow e1)$ given the experimental values obtained from the Internet domains. We employ the Sum of Square Error (SSE) as the criterion that determines the best-fitting among the following types of distributions:

- **Normal distribution** is the most commonly used distribution.
- **Cauchy distribution**: One characteristic of the Cauchy distribution is its heavy tails. In other words, it has a higher probability of extreme values.
- **Exponential distribution**: The exponential distribution is often used to model the time between events. Certain words such as *machine learning* may appear more often through time, so it becomes an excellent choice.
- **Gamma distribution**: The gamma distribution is more flexible than the exponential distribution due to the fact it has two parameters whereas the exponential distribution has one.
- **Inverse-gamma distribution**: The inverse-gamma distribution is a probability distribution of the inverse of a random variable that follows a gamma distribution.
- **Log-normal distribution**: The log-normal distribution is often used to model data that is positively skewed, but taking the logarithm of the data results a normal distribution.
- **Student's *t*-distribution**: The student's *t*-distribution is a continuous probability distribution that is similar to the normal distribution in shape but with heavier tails.

Given that we do not know the underlying distributions of word occurrences, our expectation is that if word occurrences exhibit specific characteristics, at least one of the pre-selected distributions will be able to capture distinctive features. Furthermore, all of them are often used as a prior distribution in Bayesian statistics. If a probability distribution fits the experimental values obtained from the Internet domains, simulated samples from the probability distribution should look indistinguishable compared with these experimental values. Hence, the probability distribution that has the least SSE is deemed as the best distribution. Indeed, the `fitter` package[8] returns the best distribution based on the smallest SSE and its parameters that describe the chosen distribution.

Prior Predictive Checks

After choosing a probability distribution as described in Section 4.1.1, we still need to check whether the chosen distribution is a good fit. We use Prior Predictive Checks (PPC) (Kruschke, 2015; Lambert, 2018) as a guide to judge the fit.

The concept is as follows: If we cannot tell which data is generated from the probability distribution and which one comes from the experimental values, we can conclude it is a good (enough) fit. Many statisticians use the maximum or minimum value as the criterion. In our specific case, we utilize the maximum criterion for assessing $f(e1 \rightarrow e2)$ and the minimum for evaluating $f(e2 \rightarrow e1)$. That is, it is anticipated that half of the time (i.e., 50%) the maximum or minimum value will come from simulated samples, and the other half it will come from experimental values if the chosen distribution fits best. Nevertheless, requiring an exact 50% would be overly strict, so we have extended the range to 50% $\pm$ 1% to accommodate some variability.

Algorithm 1 shows the pseudocode. M is the total number of runs that we ask the probability distribution to simulate samples; N is how many

[6] Bayesian statistics is inherently subjective in the sense that it allows individuals to express their beliefs through priors. Whether someone articulates $e1 \rightarrow e2$ or $e2 \rightarrow e1$ as expressed in Equations (12), (13), or any other forms, it remains an expression of their subjective belief.

[7] In Bayesian statistics, a probability can be interpreted as a measure of uncertainty.

[8] https://fitter.readthedocs.io/en/latest/

Algorithm 1 Prior Predictive Checks

Require: $m \geq 0, n \geq 0, i \geq 0$
$M \leftarrow m$
$c \leftarrow 0$
while $M \neq 0$ **do**
  $N \leftarrow n$
  $i \leftarrow 0$
  while $N \neq 0$ **do**
    $p \leftarrow \text{pdf}(\theta)$
    $S[i] \leftarrow p$
    $i \leftarrow i + 1$
    $N \leftarrow N - 1$
  end while
  $j \leftarrow \max(S)$ {Or $\min(S)$}
  $k \leftarrow \max(P)$ {Or $\min(P)$}
  if $j \geq k$ **then**
    $c \leftarrow c + 1$
  end if
  $M \leftarrow M - 1$
end while
return c/M

simulated samples we need for each run. Once N samples are generated, we retrieve the maximum or minimum value and store it in j. We also retrieve the maximum or minimum value from the experimental values and store it in k. If $j \geq k$ holds, we increment c by 1. Thus, c/M, which is the last line in Algorithm 1, is the percentage of times the maximum or minimum values come from simulated samples across M runs.

4.1.2 Posteriors

To approximate posterior distributions, we use the Stan open-source probabilistic programming language[9] (Kruschke, 2015; Lambert, 2018). Given Example 1, Figure 2 shows the posteriors of $f(\textit{suicide} \rightarrow \textit{death} \mid \mathbf{X})$ and $f(\textit{death} \rightarrow \textit{suicide} \mid \mathbf{X})$. These posteriors indicate that *suicide* $\rightarrow$ *death* is more likely since its posterior density is skewed toward the higher end of probabilities, making it more likely than *death* $\rightarrow$ *suicide*.

4.1.3 Bayes Factor

Given that both Model 1 and Model 2 in Equation (5) are posterior distributions, we use Bayes Factor (BF) (Lambert, 2018; McElreath, 2015) to reject either the null (i.e., Model 1) or alternative (i.e., Model 2) hypothesis. If BF is greater than 1, we opt for Model 1; Otherwise, we select Model 2.

[9] https://mc-stan.org

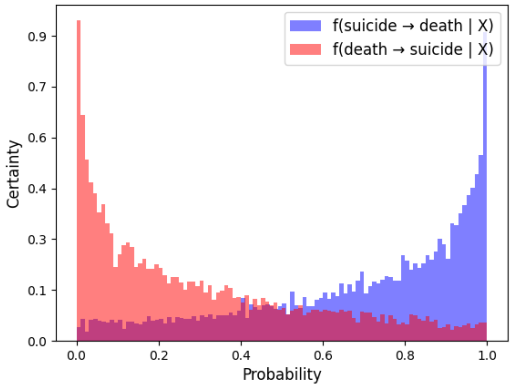


Figure 2: Both posteriors $f(\textit{death} \rightarrow \textit{suicide} \mid \mathbf{X})$ and $f(\textit{suicide} \rightarrow \textit{death} \mid \mathbf{X})$ are shown.

BF	Interpretation
BF < e^{-300}	Decisive evidence for Model 2
e^{-300} < BF < e^{300}	Reject Option (Neither)
BF > e^{300}	Decisive evidence for Model 1

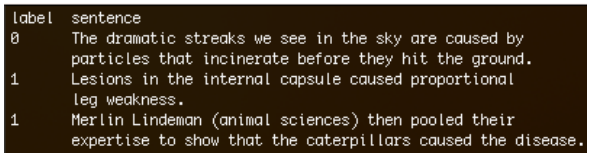
Table 3: Thresholds are used for the study.

However, it is important to note that BF tends to favour one model over the other even when both have reasonable likelihoods. Hence, Murphy (2013) suggests a threshold. By enforcing the threshold, we allow BF to make a choice only if it is confident enough. Table 3 provides a guideline about how we choose the model. When BF lies on an extreme, either a positive infinity (in which case we consider as e^{300}) or close to 0 (in which case we consider as e^{-300}), it is very confident one model is preferred over the other. Otherwise, as a Reject Option (Bishop, 2007; Murphy, 2013), neither model is chosen[10]. That is, the Bayesian framework predicts a causal direction either $e1 \rightarrow e2$, $e2 \rightarrow e1$, or *neither*. Predicted directions that fall *outside* the Reject Option will be fed to the next phase — Pre-trained BERT, as discussed in Section 4.2.

4.2 Phase 2: Pre-trained BERT

While the Bayesian framework is capable of identifying causal directions, a lack of understanding semantics means its capability is rather limited. Therefore, we turn to BERT (Devlin et al., 2019). Although BERT has many variants, we stick to the BERT uncased base model. Our implementation is largely based on a Jupyter notebook made available

[10] Strictly speaking, the Bayesian framework still predicts either $e1 \rightarrow e2$ or $e2 \rightarrow e1$ with BF falling between e^{-300} and e^{300}.



label	sentence
0	The dramatic streaks we see in the sky are caused by particles that incinerate before they hit the ground.
1	Lesions in the internal capsule caused proportional leg weakness.
1	Merlin Lindeman (animal sciences) then pooled their expertise to show that the caterpillars caused the disease.

Figure 3: The first three rows of the labelled dataset is shown.

by Rothman (2021)[11].

```
"The dramatic <e1>streaks</e1> we see in
    the sky are caused by <e2>particles
   </e2> that incinerate before they
   hit the ground."
```

Example 2: A sample sentence.

In this phase, we refine the performance of BERT, which was originally trained and made available through Hugging Face[12], by using the sentences from the test data that have causal directions predicted from the previous phase. More specifically, for each sentence in the test data, the Bayesian framework predicts either $e1 \rightarrow e2$, $e2 \rightarrow e1$, or *neither*. When the framework predicts either $e1 \rightarrow e2$ or $e2 \rightarrow e1$, we include the corresponding sentences as input to BERT, along with the predictions. Given that the Bayesian framework inherently considers uncertainty, not all sentences from the test data are passed to BERT (i.e., some have *neither*). Hence, we rely on BERT to predict those that the Bayesian framework labels *neither*. During the dataset construction process, the placeholders <e1> and <e2> are removed from sentences. `label` serves as the target variable, with 0 representing the direction $e2 \rightarrow e1$ and 1 representing the direction $e1 \rightarrow e2$. Let us take Example 2 as an example. According to the Bayesian framework, in this instance, the predicted causal direction is *particles* $\rightarrow$ *streaks* because BF < e^{-300}. Hence, we include this sentence and its predicted direction in the dataset to BERT. We continue the dataset construction process for the rest of predicted directions, as depicted in Figure 3.

5 Experiments

To evaluate our method, we have two experimental set-ups: (a) Random and (b) Bayesian + Pre-trained BERT.

[11] https://github.com/PacktPublishing/Transformers-for-Natural-Language-Processing/blob/main/Chapter02/BERT_Fine_Tuning_Sentence_Classification_DR.ipynb

[12] https://huggingface.co

To the best of our knowledge, there are no existing semi-supervised models for detecting causal directions. Thus, the random approach serves as the baseline, which blindly guesses causal directions. While one might argue that a baseline should always predict $e2 \rightarrow e1$ since it is the majority direction, it is important to note that the proposed method does not leverage such information. Hence, the random approach is more appropriate for our evaluation.

Given the SemEval-2010 (Task 8) dataset is well known, it might be tempting to consider using an established supervised model as a baseline. Using a supervised model as a baseline in a semi-supervised classification scenario is not recommended for several reasons. Firstly, supervised models are trained on labelled data whereas semi-supervised models lack annotated labels. This difference renders any experimental results incomparable: using a supervised model as a baseline can have unrealistic expectations for the performance of a semi-supervised model. Lastly, the primary objective of our study is to demonstrate the resemblance between word occurrences and the characteristics of causal directions. Using a supervised model as a baseline may distract from this objective.

(a) Random In this set-up, we simulated a probability from $Uniform(0, 1)$. If it was greater than 0.5, we would classify as $e1 \rightarrow e2$. Otherwise, $e2 \rightarrow e1$. We ran this set-up for 10,000 times and averages were recorded.

(b) Bayesian + Pre-trained BERT In this particular set-up, we ran the two-phase method described above. That is, we used the predicted directions generated from the Bayesian framework and fed them into pre-trained BERT, which made predictions on the rest of test data. This set-up was run 10 times.

We conducted the experiment under two distinct settings in the Bayesian framework. In the first setting, we examined whether the priors were satisfied with PPC (referred to as PPC+), resulting in predictions for 7 out of 328 cases. In the second setting, we did not apply any prior checks (referred to as PPC−), and this yielded predictions for 281 out of 328 cases. This allows us to gain insights into the quality of the data generated by the Bayesian framework. Because there were not enough predicted directions generated in the PPC+ setting, primarily due to a substantial number of the priors being rejected by PPC (for a detailed

Set-Up	Precision (SD[a])	Recall (SD)	F1 (SD)	Accuracy (SD)
a. Random	40.81% (2.71%)	49.99% (4.32%)	44.90% (3.18%)	49.95% (2.74%)
b. Bayesian+Pre-trained BERT (PPC+)	46.00% (2.56%)	44.93% (10.83%)	44.89% (6.22%)	55.98% (2.18%)
Bayesian+Pre-trained BERT (PPC−)	**46.82%** (1.85%)	**52.09%** (7.86%)	**49.10%** (4.14%)	**56.31%** (1.39%)

Table 4: All the experimental set-ups results are summarised.

[a]SD is short for Standard Deviation.

explanation, refer to Section 8), we augmented data by using `ContextualWordEmbsAug` from `nlpaug.augmenter` before running pre-trained BERT (Tunstall et al., 2022).

6 Results

Table 4 provides a summary of the results from all experimental set-ups (See Appendix A.1 for individual runs). In the second set-up, when the Bayesian framework creates data and feeds it into pre-trained BERT, the two-phase method yields two distinct outcomes based on the presence or absence of PPC. With Bayesian+Pre-trained BERT (PPC+), this setting achieves comparable performance to the baseline, with an F1 score of 44.89% compared to 44.90%; without PPC (i.e., PPC−), it outperforms significantly better compared with the baseline, achieving an F1 score of 49.10% versus 44.90%. To sum up, the two-phase method performs best when PPC is de-activated. PPC is necessary for assessing the trustworthiness of priors even if it led to worse performance.

7 Discussion

Although the Bayesian framework is inherently statistically sound, it is not immune to failure when confronted with certain word occurrences used in constructing the priors. In this section, we explore the Bayesian framework more comprehensively, aiming to understand the rationale behind the specific predictions made by the Bayesian framework, especially two cases from the test data where the predictions were incorrect.

1. *rain* → *cancellation* The first case, as shown in Example 3, suggests *rain* → *cancellation*, but the Bayesian framework incorrectly classified it as *cancellation* → *rain*. *rain* → *cancellation* approximates the gamma distribution whereas *cancellation* → *rain* follows the Student's *t*-distribution. Figure 4, which shows the posteriors for both *rain* → *cancellation* and *cancellation* → *rain*, clearly favours *cancellation* → *rain*.

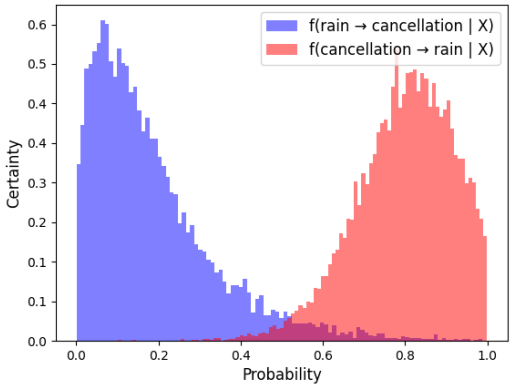


Figure 4: Both $f(rain \rightarrow cancellation \mid \mathbf{X})$ and $f(cancellation \rightarrow rain \mid \mathbf{X})$ are shown.

```
"<e1>Rain</e1> caused <e2>cancellation</
    e2> of the event in 1877, so
    enforcement of the new law had to
    wait until 1878."
Cause-Effect(e1,e2)
```

Example 3: A sample sentence. The last line indicates *rain* → *cancellation*.

Referring to Equation (10) and (11), in situations where there exists co-occurrence between $e1$ and $e2$, which is $C(e1, e2) > 0$, the entity with a higher frequency count is always identified as the effect when evaluating the entity counts. In this specific instance, the prevalence of the term $C(rain)$ typically surpasses that of $C(cancellation)$. The reason *rain* appears more often in the text could be attributed to the fact that *rain* is commonly used in everyday language, particularly weather-related contexts like events related to weather conditions.

2. *moon* → *perturbations* In the second case, as shown in Example 4, the correct answer is *moon* → *perturbations*, but the Bayesian framework erroneously misclassified it as *perturbations* → *moon*. *perturbations* → *moon* approximates the Student's *t*-distribution whereas *moon* → *perturbations* follows the inverse-gamma distribution. Figure 5, which illustrates the posterior distributions for both *moon* → *perturbations* and *perturbations* → *moon*, distinctly favours *perturbations* → *moon*.

```
"The thin F ring on the left of the
    image shows the <e1>perturbations</
    e1> caused by the <e2>moon</e2>
    Prometheus."
Cause-Effect(e2,e1)
```

Example 4: A sample sentence. The last line indicates *moon* → *perturbations*.

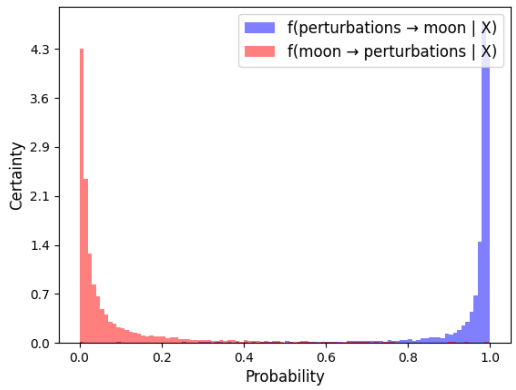


Figure 5: Both $f(\textit{moon} \rightarrow \textit{perturbations} \mid \mathbf{X})$ and $f(\textit{perturbations} \rightarrow \textit{moon} \mid \mathbf{X})$ are shown.

In consideration of Equation (10) and (11), $P(\textit{perturbations} \rightarrow \textit{moon})$ is higher than $P(\textit{moon} \rightarrow \textit{perturbations})$ in all the domains, except for wiley.com, springer.com, and ncbi.nlm.nih.gov. What they have in common is their focus on providing access to scientific research articles, publications, or resources. Given the context, which appears to be closely related to astronomy, it is likely that these specific domains cover relevant topics in this field. As further work, it is suggested to automatically identify and select the most suitable domains for the calculation of priors.

8 Further Work

There are many areas we can explore to improve the study further. In this section, we present three of them: Earth Mover's Distance, Mixture Models, and Bayesian Network.

Earth Mover's Distance. While conducting PPC in Section 4.1.1, we utilized a simple method to determine the percentage of times when the maximum or minimum value originated from simulated samples. This approach offers the advantage of being straightforward to implement because it involves comparing two numbers. However, it may not always provide reliable results. Gelman et al. (2004); Lambert (2018) recommend using Kullback-Leibler Divergence (KL Divergence) to compare two distributions. However, KL Divergence is sensitive to the choice of a reference distribution, which can be a drawback. An alternative way to do so is Earth Mover's Distance (EMD) (Rubner et al., 2000) or Word Mover's Distance (Kusner et al., 2015; Sun et al., 2019). EMD is a

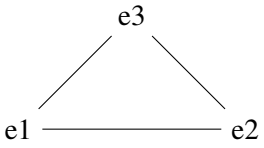


Figure 6: $e1$, $e2$ and $e3$ show causal relations.

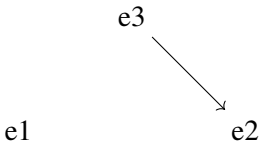


Figure 7: $e3 \rightarrow e2$ is one possible way if $e3 \rightarrow e2$ exists.

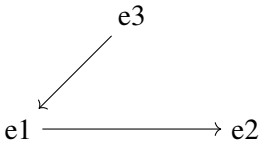


Figure 8: $e3 \rightarrow e1 \rightarrow e2$ is another possible way if $e3 \rightarrow e2$ exists.

methodology to compute "distances" between the experimental values and the listed distributions in Section 4.1.1. The distribution with the shortest distance is considered as the best fit.

Mixture Models. The distributions listed in Section 4.1.1 are not suited for modelling multi-modal data, which we frequently encountered in word occurrences, so a significant number of priors was rejected by PPC. Mixture models (Gelman et al., 2004) could be good substitutes. They are in fact probability distributions, which can account for data that exhibits multimodal and skewness. The idea is to take numerous probability distributions and stack them together using a linear combination.

Bayesian Network. We have so far considered a single causal relation in the sentence. To extend the analysis further, we can consider a multiple causal relations' scenario. That is, a model determines causal directions among all the causal relations. Let the diagram shown in Figure 6 be underlying causal relations. The task is to determine whether the causal direction $e3 \rightarrow e2$ exists. If $e3 \rightarrow e2$ exists, there are two possible networks as shown in Figures 7 and 8. We may be able to extend the proposed method to compute the likelihoods of Figures 7 and 8 if $e3 \rightarrow e2$ exists.

9 Conclusion

In this paper, we have shown empirically that word occurrences resemble the characteristics of causal directions. This finding provides significant implications and contributes significantly to our understanding of the data generation process underpinning causal directions.

Acknowledgements

We would like to thank Rolf Schwitter at Macquarie University for his advice and expertise, particularly in the area of knowledge graphs. We would also like to thank Houying Zhu at Macquarie University who offers feedback on Bayesian inference. Finally, we would like to thank Roman Marchant Matus at University of Technology Sydney for introducing us to the world of Bayesian statistics.

References

Christopher M. Bishop. 2007. *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1 edition. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*, 2nd ed. edition. Chapman and Hall/CRC.

Oren Glickman, Ido Dagan, and Moshe Koppel. 2005. Web based probabilistic textual entailment. In *Proceedings of the 1st Pascal Challenge Workshop*, pages 33–36.

J. B. S. Haldane. 1956. The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics*, 20.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Paul R. Kroeger. 2005. *Analyzing Grammar: An Introduction*. Cambridge University Press.

John Kruschke. 2015. *Doing Bayesian Data Analysis (Second Edition)*. Academic Press, Boston.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

Ben Lambert. 2018. *A Student's Guide to Bayesian Statistics*. SAGE Publications.

Zhongyang Li, Xiao Ding, Kuo Liao, Bing Qin, and Ting Liu. 2021. Causalbert: Injecting causal knowledge into pre-trained models with minimal supervision.

Richard McElreath. 2015. *Statistical Rethinking, A Course in R and Stan*. Chapman and Hall/CRC.

Farhad Moghimifar, Afshin Rahimi, Mahsa Baktashmotlagh, and Xue Li. 2020. Learning causal bayesian networks from text.

Kevin P. Murphy. 2013. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass.

Denis Rothman. 2021. *Transformers for Natural Language Processing: Build Innovative Deep Neural Network Architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and More*. Packt Publishing.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99.

Murray Singer, Michael Halldorson, Jeffrey C Lear, and Peter Andrusiak. 1992. Validation of causal bridging inferences in discourse understanding. *Journal of Memory and Language*, 31(4):507–524.

Chao Sun, King Tao Jason Ng, Philip Henville, and Roman Marchant. 2019. Hierarchical word mover distance for collaboration recommender system. In *Data Mining*, pages 289–302, Singapore. Springer Singapore.

Patrick Suppes. 1973. A probabilistic theory of causality. *British Journal for the Philosophy of Science*, 24(4):409–410.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O'Reilly Media, Incorporated.

A Appendix

A.1 Experiments

Tables 5 and 6 show the individual runs of

- Bayesian+Pre-trained BERT (PPC+), and
- Bayesian+Pre-trained BERT (PPC−)

respectively.

Run	Precision	Recall	F1	Accuracy
1	42.62%	38.81%	40.62%	53.66%
2	47.51%	64.18%	54.60%	56.40%
3	49.18%	44.78%	46.88%	58.54%
4	45.60%	42.54%	44.02%	55.79%
5	47.65%	52.99%	50.18%	57.01%
6	48.33%	43.28%	45.67%	57.93%
7	46.88%	55.97%	51.02%	56.10%
8	44.34%	35.07%	39.17%	55.49%
9	46.67%	26.12%	33.49%	57.62%
10	41.22%	45.52%	43.26%	51.22%
Average	46.00%	44.93%	44.89%	55.98%
SD	(2.56%)	(10.83%)	(6.22%)	(2.18%)

Table 5: Results of Bayesian+Pre-trained BERT (PPC+) are shown.

Run	Precision	Recall	F1	Accuracy
1	46.20%	54.48%	50.00%	55.49%
2	46.67%	57.46%	51.51%	55.79%
3	48.94%	51.49%	50.18%	58.23%
4	46.88%	55.97%	51.02%	56.10%
5	46.99%	58.21%	52.00%	56.10%
6	47.65%	52.99%	50.18%	57.01%
7	45.60%	42.54%	44.02%	55.79%
8	49.18%	44.78%	46.88%	58.54%
9	47.51%	64.18%	54.60%	56.40%
10	42.62%	38.81%	40.62%	53.66%
Average	46.82%	52.09%	49.10%	56.31%
SD	(1.85%)	(7.86%)	(4.14%)	(1.39%)

Table 6: Results of Bayesian+Pre-trained BERT (PPC−) are shown.