Automatic detection and identification of causal relationships in Polish legal texts

Łukasz Kurant[1][0000−0002−2523−5952]

University of Maria Curie-Sklodowska, Pl. M. Curie-Skłodowskiej 5, 20-031 Lublin, Poland. lukasz.kurant@mail.umcs.pl

Abstract. The paper focuses on the problem of detecting sentences containing causal relations in Polish legal texts. The identification of these relationships and their decomposition is a key factor in the effective analysis of legal texts and an important aspect in the extraction of parts of such relationships. This represents a contribution to the development of the field for languages other than English. The paper presents an analysis of the created dataset and based on it, classification was performed in nine different experiments using selected machine learning and deep learning algorithms (including several large BART-type models), taking into account the specifics of legal language. The experiments confirm the effectiveness of the proposed method, where the best model detected sentences containing both explicit and implicit causality with an accuracy of approximately 86%. These results lead to further questions and point to further directions for future development, especially in the field of reasoning from legal texts.

Keywords: causal relationships · argumentation · artificial intelligence.

1 Introduction

Detecting cause-and-effect relationships is a challenge in natural language processing. This requires advanced cognitive processes, and the resulting data has wide applications in many scientific fields. However, there are several challenges to overcome [4], such as minimizing ambiguity and recognizing causal relationships that can exist in both explicit and implicit forms. Detecting relationships can be a significant obstacle, especially in the latter case. As with other Natural Language Processing (NLP) tasks, it is important to consider the impact of natural language, including its structural and semantic aspects such as vocabulary, sarcasm, and metaphors. Both the language itself and the specific domain can have an impact. Research on the influence of language on the process of formulating cause and effect should not be limited to specific domains or languages. Therefore, it is important to consider a wide range research in this subject. Detected causal relationships can be widely applied in many fields [2], including the field of law. Accurate causal reasoning is crucial for legal professionals in their daily work. For instance, judges use it to formulate sentences based on similar case law, while attorneys and prosecutors use it to determine the appropriate courtroom strategy.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15

Two main tasks can be distinguished in causality detection. The first task is to identify the locations (e.g. sentences) where causality occurs. Based on this information, the parts of the relationship can be extracted, determining the cause and effect, and filtering out irrelevant information. It is important to maintain a clear and logical flow of information with causal connections between statements.

2 Definitions

We can define causality as the relationship between two different events $event_1$ and $event_2$ in such a way that $event_2$ results from $event_1$. Various approaches are used to formalize the definition of causality, e.g. as an implication where the occurrence of cause c_1 implies the occurrence of effect e_1 ($c_1 \Rightarrow e_1$) or as a logical equivalent ($c_1 \Leftrightarrow e_2$), for the reason of avoiding ambiguity [12]. The choice of definition may thus depend on the specific problem that practitioners were to solve [3]. Therefore, due to the fact that during our experiments it does not matter whether an effect can occur for reasons other than those written out, we decided to use logical equivalence notation.

2.1 Division by type

According to the definition of the type of relationship shown in [4], causal relationships can be divided into three categories as follows:

- $c_1 \Leftrightarrow e_1$, if c_1 exists then e_1 also exists, e.g. "The judge convicted him because the evidence was against him".
- $c_1 \Leftrightarrow \neg e_1$, if c_1 exists then e_1 does not exist, e.g. "It is clear from the witness's testimony that he could not do so".
- $\neg c_1 \Leftrightarrow \neg e_1$, if c_1 does not exist then e_1 does not exist either, e.g. "The witness was not allowed into the courtroom because he did not have a valid identity card".

2.2 Division by complexity

We can also divide causality according to other criteria. Causality can be single-sentence or multi-sentence. In the case of the latter, the process of detection or extraction is much more complicated [25]. In some cases, causality may be more complex, such that the number of causes and effects may not be equal, i.e. many different causes may cause one effect, or vice versa: one cause may lead to many effects. Both causes and effects can be connected by conjunctions ($c_1 \Leftrightarrow e_1 \wedge e_2$) or by disjunctions, e.g. ($c_1 \Leftrightarrow e_1 \vee e_2$) in any way, including as combinations of these two. Because of this, causality can also lead to so-called causal chains, in the way that the effect of a cause, can be the cause of another effect, e.g. "The court did not allow a witness into the courtroom, due to an invalid identity document, which resulted in the person against whom the proceedings were taking place being found not guilty". We can therefore define such a chain as: "invalid document (c_1) $\Leftrightarrow$ witness not allowed (e_1/c_2) $\Leftrightarrow$ not guilty (e_2)".

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15

2.3 Division by form

We can also divide causality based on its form:

- *explicit causality*, occurring overtly in sentences, often with phrases or conjunctions indicating causality, e.g. "The judge convicted him because the evidence was against him",
- *implicit causality*, occurring implicitly, often with parts divided between different sentences, e.g. "The judge found him not guilty. No evidence of his guilt was found".

Within explicit causality, we can further distinguish: *marked causality*, when the text contains causal conjunctions, such as "because", "as"; and *unmarked causality*, when the text does not contain such a conjunct, but contains other phrases (e.g. verbs) that indicate the existence of causality. In addition, each linker, causal phrase can be divided into one of two categories, i.e. *ambiguous causal phrase*, when a phrase or word can be used in different contexts and only in some of them its use proves causality, and *non-ambiguous causal phrase*, when causality is inferred from almost every instance of the phrase's use.

2.4 Division by order of occurrence

The components of cause-effect relationships can be related to each other in different ways based on the timing of their occurrence. There are many different divisions based on such criteria, e.g. TimeML [18], which distinguishes as many as seven different types of temporal relationships, or CaTeRS [13], which distinguishes four types of them. Due to the complexity of the relationships of these types, we can simplify them into two groups:

- the cause occurred before the effect, e.g. "The witness failed to attend due to a car breakdown",
- the cause occurred together with the effect, e.g. "The witness spoke slowly because he had a speech defect".

In the second case, it does not matter whether the cause and effect occurred together throughout their existence – if there was a concurrent time, the relation is qualified in this category.

3 Research status

Among the methods currently used in the field of causality detection, we can distinguish between rule-based and pattern-based methods [6], statistical methods (including machine learning methods such as linear regression, Naive Bayes or decision trees), and more sophisticated methods using deep learning and neural networks. Among the latter, we can distinguish architectures such as CNNs, LSTMs, GRUs and Transformers such as BERT [4]. Each of these methods has its own advantages and disadvantages. Methods using patterns require domain

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15

expertise, while methods using machine learning need to be programmed and trained, which can result in the need for a large amount of resources. When detecting causality, the use of word embeddings also plays an important role and has a significant impact on improving performance [5].

The main problem during research is the lack of sufficient datasets, especially in the context of languages other than English. The rudimentary yet underdeveloped notion of causality in Polish legat texts was introduced in the [9]. In addition, the domain from which the samples in the collection are drawn is important. In this case, it is difficult to compare the chosen methods, due to the specificities of the language, in which certain features may be useless [8]. In particular, it is a difficult task to prepare a collection containing numerous instances of implicit causation, due to the difficulty for annotators to recognize it.

The process of detecting whether a relationship exists in a text is often the first step to extracting the components of a given causal relationship, but it is an important part of it. Hence, it is significant to get the best possible results in the first step, in order to avoid potential cascading errors in the future (when a sentence without causality is flagged by the model as having such a relationship, this can lead to further extraction errors [11]).

4 Dataset

The aim of our experiment was to perform causality detection at the sentence level. For this, it became necessary to prepare a suitable dataset, which would take into account not only whether a given sentence contains a cause-effect relation, but also the type of this relation, its complexity or form. For this purpose, 50 different court judgments in Polish were selected, from five different categories (10 judgments from each): animal protection, taxes, juvenile, infringement of privacy, international law.

The court judgments are taken from the publicly available Portal of Administrative Court Judgments [23]. Due to their nature, they have anonymised sensitive data, such as the names of individuals, place names, etc. Using the author's script (using the *beautifulsoup* library in Python [20]), the sentences were downloaded in HTML format and appropriately processed to split them into sentences (using the *NLTK* library [16]). The sentences prepared in this way were then imported into the *doccano* software [15] used to prepare their annotation. During the annotation process, the following tags were added for each sentence:

- *Causality / No Causality* – if the sentence contains a causality relationship,
- *Not valid sentence* – whether the sentence is a valid sentence. Legal texts sometimes contain referenced provisions or other phrases that should not be included in the detection process.

If causality exists then:

- *Implicit causality / Explicit causality* – whether causality is explicit or non-explicit,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15

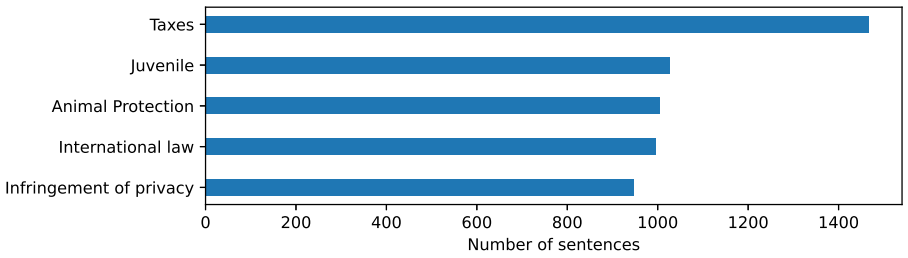

Fig. 1. Number of sentences by category of court judgement

- *Single cause* / *Multiple causes* – when there is only one, or many different causes,
- *Single effect* / *Multiple effects* – when there is only one, or many different effects,
- *Event chain* – if sentence contains chain of causality,
- $c_1 \Leftrightarrow e_1$ *relation* / $c_1 \Leftrightarrow \neg e_1$ *relation* / $\neg c_1 \Leftrightarrow \neg e_1$ *relation* – relation by type of constituent parts,
- *Cause before effect* / *cause together with effect* – according to the timing of cause and effect.

If explicit causality exists then:

- *Marked* / *Unmarked* – if the sentence contains phrase suggesting causality,
- *Ambiguous causal phrase* / *non ambiguous causal phrase* – if a phrase suggesting causality does so only depending on the context or almost always.

The annotation process took place in two stages. First, we annotated the entire dataset, then we performed verification on the same dataset, but without access to the previously added tags. By comparing sentences whose tags differed between the two processes, we only added those that were more appropriate. In the process of preparing the collection for annotation, all the anonymised data referred to above were replaced by the tokens "ENTITY_0", "ENTITY_1", etc. depending on the number of unique abbreviations present in the sentence.

The sentences, which are the components of the judgments, have a specific, very formal structure in which words and phrases specific to the legal language are found (including often possessing phrases containing Latin). The detailed number of sentences in each category is shown in Figure 1. As we can infer from it, this set is largely unbalanced in terms of the number of sentences with a causal link to those without. Nevertheless, this provides an indication that sentences that contain causality comprise a large proportion of all sentences. Among the items in the test set, sentences with a character count between 100 and 200 were the largest group (Figure 2). Due to the fact that legal texts often contain very elaborate sentences, there is also a large representation of samples with a much higher number of characters. In each type of category, there is a certain number of sentences with causality, without it, and those that are not correct sentences,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15

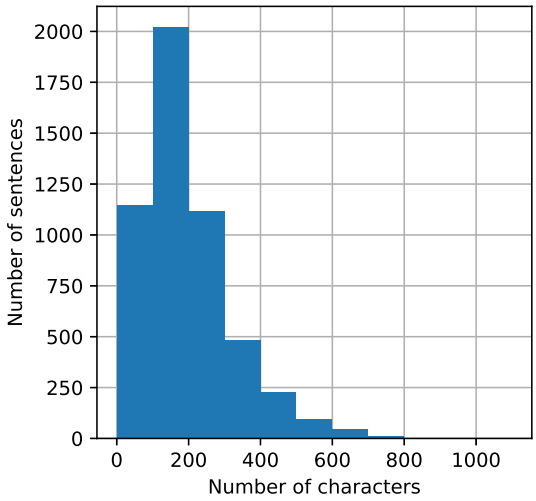

Fig. 2. Number of sentences with the selected character range

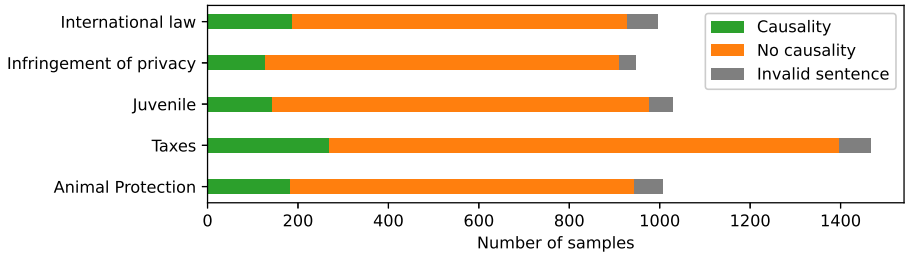

Fig. 3. Number of samples with Causality, No causality and Invalid sentence tags by category

so they will not be analysed further (such sentences in the entire set are about 5.25%). A detailed division by category is shown in Figure 3.

In the dataset, each sentence was assigned appropriate tags as described above. The detailed number of elements of a given class, together with the percentage, is shown in Figure 4. The collection for later use has been exported to JSON format, which is the input for the programmes using the selected machine learning models described in the next chapter.

Sentences with explicit causality, as described above, often have words or phrases that suggest the existence of such a relationship. Figure 5 shows a list of the most common words found in this type of sentence. These words, such as “gdyż” (because), “wynika” (follows) or “jeżeli” (if) are typical phrases from which causality is implied (other words, popular for a specific variety of language are also found).

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15

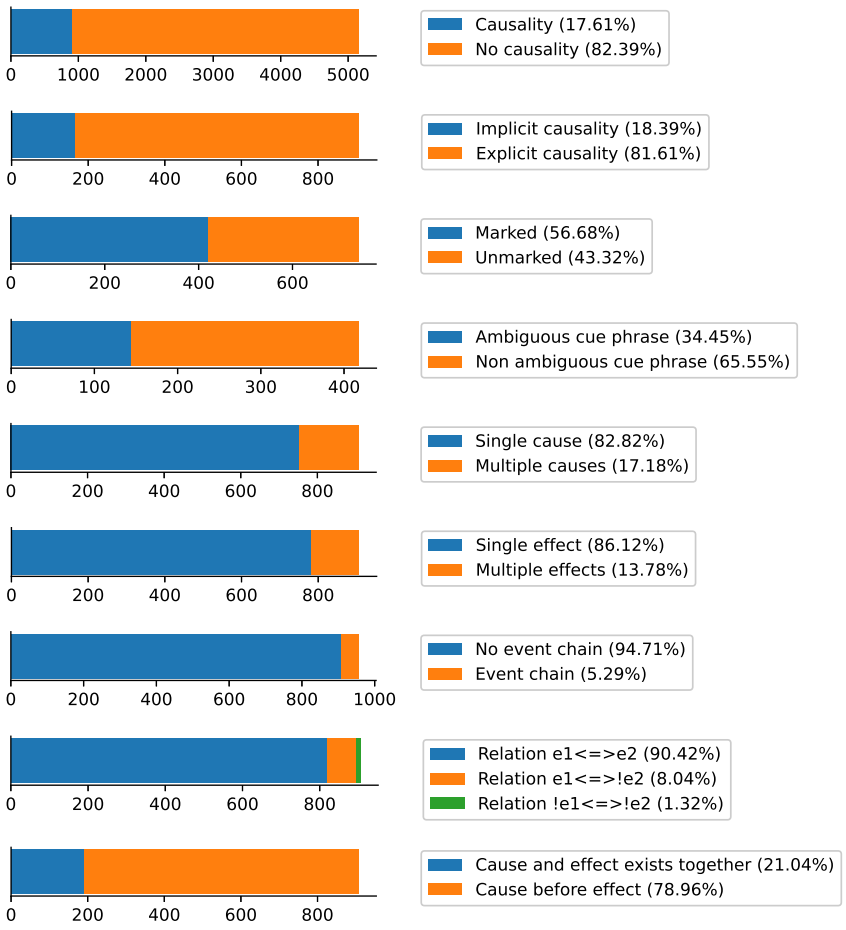

Fig. 4. Division of samples by class

5 Experiments

The main goal of the experiment was to detect sentences with a causal relationship, with a distinction between implicit and explicit causality. For this purpose, different classification methods were prepared and after a training process, they were appropriately evaluated using standard metrics such as Accuracy (ACC), Precision (P), Recall (R) and F1 measure. Classifiers from two categories were used for this purpose: machine learning-based and neural network-based. In each case, we dealt with binary classification, in the following classes (the first class is treated as positive and the second as negative):

– **(E1)** Causality / No Causality,
– **(E2)** Implicit causality / Explicit causality,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15

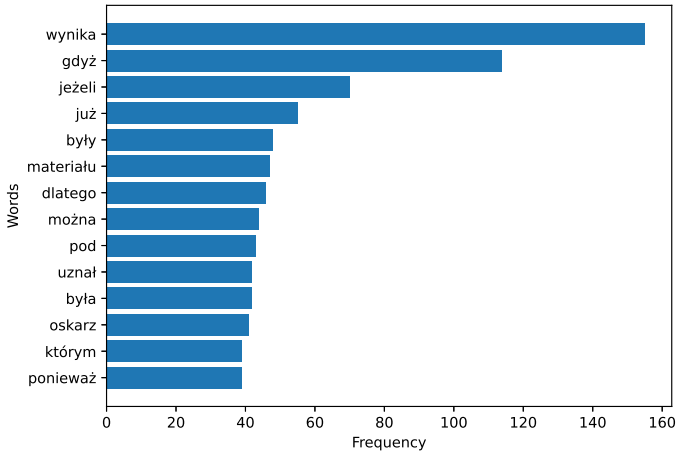

Fig. 5. The most popular words along with the number of occurrences in sentences containing causality relation.

- **(E3)** Marked / Unmarked,
- **(E4)** Ambiguous causal phrase / non ambiguous causal phrase,
- **(E6)** Single cause / Multiple causes,
- **(E6)** Single effect / Multiple effects,
- **(E7)** Event chain / No event chain,
- **(E8)** $c_1 \Leftrightarrow e_1$ relation / $c_1 \Leftrightarrow \neg e_1$, relation (due to insufficient number of samples, the class $\neg c_1 \Leftrightarrow \neg e_1$ was omitted),
- **(E9)** Cause before effect / cause together with effect.

Nine different experiments were therefore conducted, each with 16 different models. Because the sets were not balanced, to equalize the number of elements of the classes during training, the number of samples per class was equalized to the number of elements from the least numerous class during a given experiment. The equalization involved selecting random values from the dataset using default mechanisms contained in the Scikit-learn library[17].

5.1 Machine learning-based methods

The first type of classifiers were machine learning models trained in a supervised manner based on a labeled dataset. The following models were used: Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Random Forest (RF) and XGBoost (XGB). The data were prepared in three formats: Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF) and in the form of dense vectors (word embeddings, WE) – pre-prepared FestText vector sets for Polish, trained using the CBOW technique, stored in 300 dimensions using n-grams of characters of length 5 and windows size equals 10 negatives [7]. A detailed list of hyperparameters is presented in Table 1.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15

Table 1. Hyperparameters of the models used in the experiments

Model	Input	Hyperparameters
MNB	BoW, TF-IDF	alpha: 1, fit_prior: True
SVM	BoW, TF-IDF	c: 1, kernel: rbf, degree: 3, gamma: scale, tol: 1e-3
KNN	BoW	n neighbors: 20, algorithm: ball_tree, leaf_size: 30
RF	BoW	n estimators: 100, criterion: gini, max features: sqrt
XGB	BoW, WE	n estimators: 1000, subsample: 0.8, early stopping rounds: 10
SN	WE	optimizer: adam, loss: Binary Crossentropy
CNN	WE	optimizer: RMSprop (lr: 1e-4), loss: Binary Crossentropy
BiLSTM	WE	optimizer: adam, loss: Binary Crossentropy, lstm units: 32
Transformer	WE	embed_dim: 32, num_heads: 2, ff_dim: 32, maxlen: 300, optimizer: adam, learning_rate: 1e-4, loss: Binary Crossentropy
DistilBERT	BERT tokens	104 languages, embed_dim: 768, hidden_layers: 7, num_heads: 12
RoBERTa	BERT tokens	Polish only, embed_dim: 768, hidden_layers: 12, num_heads: 12
HerBERT	BERT tokens	Polish only, embed_dim: 768, hidden_layers: 12, num_heads: 12
Polbert	BERT tokens	Polish only, embed_dim: 768, hidden_layers: 12, num_heads: 12

5.2 Neural network-based methods

The second type of classifiers are methods based on selected neural networks. Several popular architectures based on such networks were chosen for the experiment, such as: Shallow Neural Network (SN, only with one hidden layer), Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory Network (BiLSTM), Transformer Network (TN). In each of these cases, the input data had the format of word embeddings vectors (the same as described in the section on machine learning methods).

In addition, other models based on Bidirectional Encoder Representations from Transformers (BERT), prepared for the Polish language on which the fine-tuning process was carried out, were also used: DistilBERT (base, multilingual, cased) [22], Polish RoBERTa v2 (large) [1], HerBERT (base, cased) [14], Polbert (base, uncased) [10]. All these models were used with the *Simple Transformers* library [19]. The models were selected based on the results of the KLEJ benchmark [21] (the GLUE equivalent for English models [24]). As for the previous models, the hyperparameters of the models are detailed in Table 1.

Causal words or phrases are a great help in identifying obvious causation. In the case of legal texts, such words differ to some extent from words used in informal language. For example, the word "albowiem" (because, but very formal) may or may not be used in causal contexts. Models can also fail when a sen-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15

Table 2. Results of the Experiment 1 (E1)

Model		Macro			Causality			No causality		
	AC	P	R	F1	P	R	F1	P	R	F1
NB (BoW)	0.68	0.71	0.68	0.67	0.63	0.86	0.73	0.78	0.51	0.61
NB (TF-IDF)	0.66	0.69	0.66	0.64	0.61	0.87	0.72	0.78	0.44	0.56
SVM (BoW)	0.74	0.74	0.74	0.74	0.74	0.73	0.73	0.73	0.75	0.74
SVM (TF-IDF)	0.75	0.75	0.75	0.75	0.78	0.69	0.73	0.72	0.80	0.76
KNN	0.54	0.71	0.54	0.43	**0.91**	0.10	0.17	0.52	**0.99**	0.68
RF	0.77	0.77	0.77	0.77	0.81	0.71	0.76	0.74	0.83	0.78
XGB (BoW)	0.79	0.79	0.79	0.78	0.83	0.72	0.77	0.75	0.85	0.80
XGB (WE)	0.85	0.86	0.85	0.85	0.81	**0.93**	0.86	**0.91**	0.78	0.84
SN	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.74	0.73
CNN	0.70	0.74	0.70	0.69	0.65	0.89	0.75	0.82	0.51	0.63
BiLSTM	0.73	0.73	0.73	0.73	0.74	0.71	0.72	0.72	0.75	0.74
TN	0.72	0.74	0.72	0.72	0.68	0.85	0.75	0.79	0.59	0.68
DistilBERT	0.82	0.82	0.82	0.82	0.80	0.85	0.82	0.84	0.79	0.81
RoBERTa	**0.86**	**0.87**	**0.86**	**0.86**	0.83	0.92	**0.87**	0.91	0.81	**0.86**
HerBERT	0.85	0.85	0.85	0.85	0.81	0.91	0.86	0.89	0.79	0.84
Polbert	0.79	0.79	0.79	0.78	0.77	0.82	0.79	0.80	0.75	0.78

tence contains a cause-effect relationship, but either the cause or effect is split across multiple sentences. This can result in a lack of relationship at the sentence level, highlighting the need to study such links at a larger level than just one sentence. Considering the complex sentence structures commonly found in legal documents, it is important to keep in mind that these constructions can often be lengthy due to the presence of multiple subordinate clauses.

6 Results

The main experiment (E1) was the detection of sentences with cause-and-effect relationships. Table 2 shows the detailed results for this task by model. Depending on the model, accuracy scores range from 0.54 to 0.87. The best performer was the RoBERTa model, which presented good results for both the class with and without causality. Similar results were obtained by the XGBoost model based on word embeddings vectors. Detecting causality in legal texts is therefore a possible task, although quite difficult in the case of implicit causality. In the case of the latter experiment (E2), the results are also solid, but also in this case, the RoBERTa model was the best, achieving an accuracy of 0.90. In the results, we can also notice a regularity that models based on BERT trained using only Polish texts perform noticeably better than the multilingual model. In the case of models based on machine learning, there is not much difference between the results of models based on BoW or TF-IDF. When evaluating such models, we should keep in mind that if we wanted to use this type of data in further experiments (e.g. extraction of such compounds), we should focus on better Recall results than Precision, due to the cascade errors described earlier. The

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15

Table 3. Summary results of accuracy in experiments E2-E9

Model	Accuracy							
	E2	E3	E4	E5	E6	E7	E8	E9
NB (BoW)	0.65	0.67	0.65	0.64	0.62	0.66	0.64	0.65
NB (TF-IDF)	0.66	0.65	0.61	0.63	0.60	0.66	0.64	0.65
SVM (BoW)	0.77	0.80	0.75	0.74	0.71	0.88	0.73	0.72
SVM (TF-IDF)	0.76	0.77	0.77	0.71	0.70	0.78	0.72	0.73
KNN	0.52	0.54	0.52	0.54	0.53	0.56	0.53	0.50
RF	0.78	0.83	0.80	0.75	0.74	0.88	0.75	0.75
XGB (BoW)	0.83	**0.93**	**0.90**	0.76	0.75	**0.91**	0.76	0.72
XGB (WE)	0.84	0.84	0.89	0.80	0.79	0.91	0.83	**0.83**
SN	0.76	0.77	0.50	0.71	0.68	0.50	0.73	0.50
CNN	0.70	0.65	0.61	0.67	0.68	0.88	0.72	0.63
BiLSTM	0.73	0.77	0.50	0.74	0.68	0.50	0.72	0.50
TN	0.68	0.75	0.50	0.71	0.56	0.50	0.71	0.50
DistilBERT	0.85	0.84	0.72	0.76	0.80	0.72	0.84	0.65
RoBERTa	**0.90**	0.91	0.84	**0.86**	**0.86**	0.84	**0.88**	0.76
HerBERT	0.86	0.91	0.81	0.85	0.81	0.56	0.85	0.65
Polbert	0.80	0.78	0.80	0.74	0.79	0.75	0.80	0.65

situation is similar for the other experiments (E3-9). The results for accuracy for the experiments are shown in Table 3. Detailed results for both classes are shown in Table 4.

7 Conclusions

The main issue with this type of analysis, in terms of causality, is the absence of a suitable dataset, particularly for languages other than English. Therefore, it is essential to create such a collection yourself, which can be a time-consuming task, given the chosen domain of texts, such as legal texts.

Excluding legal texts from analysis can facilitate the process, but it can also create issues when attempting to generalize methods. It is important to note that this collection contains various types of documents, including statutory texts and court judgments, which may differ significantly in their use of formal language. However, court judgments often share a similar structure, typically including the same components, such as the operative part, grounds, and referenced provisions.

Upon analysis of the collection, it can be concluded that although causality sentences make up less than 20% of the collection, they convey crucial information for future legal analysis. Additionally, determining the type of causality can provide significant information, particularly in extracting the constituent parts of such compounds. Explicit causality is the main form in which causality occurs. However, detecting implicit causality can be challenging due to the ambiguity involved. Our empirical findings confirm that detecting causality in sentences from real-world data, such as court judgments, can be achieved with a satisfactory F1 index score of 86%.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15

8 Future work

After a thorough analysis of the relationships present in legal texts, it is worth considering the extraction of the components of these relationships. The extracted parts, including causes and effects, can provide important data for further analysis, however, it is necessary to construct a specific models to identify these components. The prevailing view in current research is that it is important to develop general methods that can be applied to different fields. However, it is worth noting that in some fields, such as medicine or law, the languages used are so specific that they can pose a significant challenge for this type of analysis. It is therefore crucial to carry out research dedicated to specific fields. Another important step is to identify causality relations between a wider spectrum of sentences. Often sentences in close neighborhood exhibit relationships that are not discernible through single-sentence analysis. Models using the attention mechanism can therefore be used for this purpose, in such a way as to find connections between fragments of text located further apart.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Dadas, S., Perełkiewicz, M., Poświata, R.: Pre-training polish transformer-based language models at scale. In: Artificial Intelligence and Soft Computing. pp. 301–314. Springer International Publishing (2020). https://doi.org/10.48550/arXiv.2006.04229
2. Dasgupta, T., Saha, R., Dey, L., Naskar, A.: Automatic extraction of causal relations from text using linguistically informed deep neural networks. In: Komatani, K., Litman, D., Yu, K., Papangelis, A., Cavedon, L., Nakano, M. (eds.) Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue. pp. 306–316. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/W18-5035
3. Fischbach, J., Frattini, J., Méndez, D., Unterkalmsteiner, M., Femmer, H., Vogelsang, A.: How do practitioners interpret conditionals in requirements? CoRR **abs/2109.02063** (2021). https://doi.org/10.48550/arXiv.2109.02063
4. Frattini, J., Fischbach, J., Mendez, D., Unterkalmsteiner, M., Vogelsang, A., Wnuk, K.: Causality in requirements artifacts: prevalence, detection, and impact. Requirements Engineering **28**(1), 49–74 (Mar 2023). https://doi.org/10.1007/s00766-022-00371-x
5. Girju, R.: Automatic detection of causal relations for question answering. In: Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering. pp. 76–83. Association for Computational Linguistics, Sapporo, Japan (Jul 2003). https://doi.org/10.3115/1119312.1119322
6. Girju, R., Moldovan, D.: Text mining for causal relations. In: Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference. pp. 360–364. The Florida AI Research Society (10 2002)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15

7. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018) (2018). https://doi.org/10.48550/arXiv.1802.06893
8. Keskes, I., Zitoune, F.B., Belguith, L.H.: Learning explicit and implicit arabic discourse relations. Journal of King Saud University - Computer and Information Sciences **26**(4), 398–416 (2014). https://doi.org/https://doi.org/10.1016/j.jksuci.2014.06.001
9. Kurant, L.: Mechanism for detecting cause-and-effect relationships in court judgments. In: Proceedings of the 18th Conference on Computer Science and Intelligence Systems. vol. 35, p. 1041–1046. ACSIS (2023). https://doi.org/10.15439/2023F4827
10. Kłeczek, D.: Polbert - polish bert (2020), https://huggingface.co/dkleczek/bert-base-polish-uncased-v1
11. Li, Z., Li, Q., Zou, X., Ren, J.: Causality extraction based on self-attentive bilstm-crf with transferred embeddings. CoRR **abs/1904.07629** (2019), http://arxiv.org/abs/1904.07629
12. Mavin, A., Wilkinson, P., Harwood, A., Novak, M.: Easy approach to requirements syntax (ears). In: Requirements Engineering Conference, 2009. RE '09. 17th IEEE International. pp. 317 – 322 (10 2009). https://doi.org/10.1109/RE.2009.9
13. Mostafazadeh, N., Grealish, A., Chambers, N., Allen, J., Vanderwende, L.: CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In: Palmer, M., Hovy, E., Mitamura, T., O'Gorman, T. (eds.) Proceedings of the Fourth Workshop on Events. pp. 51–61. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/W16-1007
14. Mroczkowski, R., Rybak, P., Wróblewska, A., Gawlik, I.: HerBERT: Efficiently pretrained transformer-based language model for Polish. In: Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing. pp. 1–10. Association for Computational Linguistics, Kiyv, Ukraine (Apr 2021), https://www.aclweb.org/anthology/2021.bsnlp-1.1
15. Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., Liang, X.: doccano: Text annotation tool for human (2018), https://github.com/doccano/doccano
16. NLTK: Natural language toolkit (2023), https://www.nltk.org, accessed: 2024-01-12
17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
18. Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., Katz, G., Radev, D.: Timeml: Robust specification of event and temporal expressions in text. pp. 28–34 (01 2003)
19. Rajapakse, T.C.: Simple transformers. https://github.com/ThilinaRajapakse/simpletransformers (2019), accessed: 2024-01-12
20. Richardson, L.: Beautiful soup python library (2004-2023), https://www.crummy.com/software/BeautifulSoup/bs4/doc/, accessed: 2024-01-12
21. Rybak, P., Mroczkowski, R., Tracz, J., Gawlik, I.: KLEJ: comprehensive benchmark for polish language understanding. CoRR **abs/2005.00630** (2020). https://doi.org/10.48550/arXiv.2005.00630

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15

22. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv **abs/1910.01108** (2019). https://doi.org/10.48550/arXiv.1910.01108
23. Sprawiedliwości, M.: Portal orzeczeń sądów powszechnych (2012-2024), https://orzeczenia.ms.gov.pl, accessed: 2023-11-04
24. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding (2018)
25. Yang, J., Han, S.C., Poon, J.: A survey on extraction of causal relations from natural language text. CoRR **abs/2101.06426** (2021). https://doi.org/10.48550/arXiv.2101.06426

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15

Table 4. Detailed results of experiments E2-E9

Experiment 2	Macro			Positive			Negative		
	P	R	F1	P	R	F1	P	R	F1
NB (BoW)	0.70	0.65	0.62	0.60	0.91	0.72	0.81	0.39	0.52
NB (TF-IDF)	0.73	0.66	0.63	0.60	0.93	0.73	0.85	0.38	0.52
SVM (BoW)	0.77	0.77	0.77	0.77	0.78	0.77	0.77	0.76	0.77
SVM (TF-IDF)	0.76	0.76	0.76	0.77	0.75	0.76	0.75	0.78	0.77
KNN	0.72	0.52	0.39	**0.93**	0.05	0.10	0.51	**1.00**	0.68
RF	0.78	0.78	0.78	0.80	0.75	0.77	0.77	0.81	0.79
XGB (BoW)	0.84	0.83	0.83	0.85	0.82	0.83	0.82	0.85	0.84
XGB (WE)	0.85	0.84	0.84	0.81	0.91	0.85	0.90	0.78	0.83
SN	0.76	0.76	0.76	0.76	0.75	0.76	0.76	0.77	0.76
CNN	0.74	0.70	0.68	0.64	0.92	0.75	0.85	0.47	0.61
BiLSTM	0.77	0.73	0.73	0.67	0.91	0.77	0.86	0.56	0.68
TN	0.71	0.68	0.66	0.79	0.48	0.60	0.63	0.87	0.73
DistilBERT	0.85	0.85	0.85	0.83	0.88	0.86	0.87	0.82	0.85
RoBERTa	**0.91**	**0.90**	**0.90**	0.86	**0.96**	**0.91**	**0.96**	0.84	**0.89**
HerBERT	0.87	0.86	0.86	0.80	0.96	0.87	0.94	0.77	0.85
Polbert	0.80	0.80	0.80	0.78	0.84	0.81	0.83	0.76	0.79

Experiment 3	Macro			Positive			Negative		
	P	R	F1	P	R	F1	P	R	F1
NB (BoW)	0.70	0.67	0.66	0.62	0.86	0.72	0.77	0.48	0.59
NB (TF-IDF)	0.70	0.65	0.63	0.60	0.90	0.72	0.80	0.41	0.54
SVM (BoW)	0.80	0.80	0.80	0.78	0.83	0.80	0.82	0.77	0.79
SVM (TF-IDF)	0.77	0.77	0.77	0.78	0.77	0.77	0.77	0.78	0.77
KNN	0.73	0.54	0.43	**0.93**	0.09	0.17	0.52	**0.99**	0.68
RF	0.83	0.83	0.83	0.82	0.84	0.83	0.84	0.82	0.83
XGB (BoW)	**0.93**	**0.93**	**0.93**	0.93	0.93	**0.93**	0.93	0.93	**0.93**
XGB (WE)	0.86	0.84	0.84	0.78	0.94	0.86	0.93	0.74	0.82
SN	0.77	0.77	0.77	0.78	0.76	0.77	0.77	0.78	0.78
CNN	0.76	0.65	0.60	0.59	**0.98**	0.74	0.94	0.32	0.47
BiLSTM	0.77	0.77	0.77	0.81	0.71	0.75	0.74	0.83	0.78
TN	0.75	0.75	0.75	0.75	0.76	0.75	0.75	0.75	0.75
DistilBERT	0.84	0.84	0.84	0.82	0.87	0.85	0.86	0.81	0.84
RoBERTa	0.91	0.91	0.91	0.87	0.96	0.91	**0.96**	0.86	0.90
HerBERT	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
Polbert	0.78	0.78	0.78	0.80	0.74	0.77	0.76	0.82	0.79

Experiment 4	Macro			Positive			Negative		
	P	R	F1	P	R	F1	P	R	F1
NB (BoW)	0.68	0.65	0.63	0.60	0.85	0.71	0.75	0.44	0.55
NB (TF-IDF)	0.66	0.61	0.59	0.58	0.88	0.69	0.74	0.35	0.48
SVM (BoW)	0.76	0.75	0.75	0.71	0.83	0.77	0.80	0.67	0.73
SVM (TF-IDF)	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77
KNN	0.76	0.52	0.38	**1.00**	0.04	0.08	0.51	**1.00**	0.68
RF	0.81	0.80	0.80	0.84	0.75	0.79	0.77	0.85	0.81
XGB (BoW)	**0.90**	**0.90**	**0.90**	0.93	0.85	**0.89**	0.87	0.94	**0.90**
XGB (WE)	0.89	0.89	0.89	0.86	0.92	0.89	0.91	0.85	0.88
SN	0.25	0.50	0.33	0.50	**1.00**	0.67	0.00	0.00	0.00
CNN	0.78	0.61	0.55	0.56	1.00	0.72	**1.00**	0.23	0.37
BiLSTM	0.25	0.50	0.33	0.50	1.00	0.67	0.00	0.00	0.00
TN	0.25	0.50	0.33	0.50	1.00	0.67	0.00	0.00	0.00
DistilBERT	0.72	0.72	0.72	0.72	0.71	0.72	0.71	0.73	0.72
RoBERTa	0.85	0.84	0.84	0.81	0.90	0.85	0.88	0.79	0.84
HerBERT	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
Polbert	0.80	0.80	0.80	0.81	0.79	0.80	0.80	0.81	0.80

Experiment 5	Macro			Positive			Negative		
	P	R	F1	P	R	F1	P	R	F1
NB (BoW)	0.66	0.64	0.64	0.71	0.49	0.58	0.61	0.80	0.69
NB (TF-IDF)	0.66	0.63	0.61	0.73	0.41	0.52	0.59	**0.85**	0.69
SVM (BoW)	0.74	0.74	0.74	0.76	0.71	0.74	0.73	0.77	0.75
SVM (TF-IDF)	0.71	0.71	0.71	0.71	0.73	0.72	0.72	0.70	0.71
KNN	0.76	0.53	0.41	0.52	**1.00**	0.68	**1.00**	0.07	0.13
RF	0.75	0.75	0.75	0.72	0.83	0.77	0.79	0.67	0.73
XGB (BoW)	0.78	0.76	0.76	0.72	0.88	0.79	0.84	0.65	0.73
XGB (WE)	0.80	0.80	0.80	0.78	0.84	0.81	0.82	0.76	0.79
SN	0.71	0.71	0.71	0.69	0.78	0.73	0.74	0.64	0.69
CNN	0.68	0.67	0.66	0.74	0.53	0.61	0.63	0.81	0.71
BiLSTM	0.75	0.74	0.74	0.78	0.67	0.72	0.71	0.81	0.76
TN	0.71	0.71	0.71	0.72	0.71	0.71	0.71	0.72	0.71
DistilBERT	0.76	0.76	0.76	0.74	0.79	0.77	0.78	0.72	0.75
RoBERTa	**0.87**	**0.86**	**0.86**	**0.84**	0.90	**0.87**	0.90	0.82	**0.86**
HerBERT	0.85	0.85	0.84	0.83	0.88	0.85	0.87	0.82	0.84
Polbert	0.74	0.74	0.74	0.74	0.76	0.75	0.75	0.73	0.74

Experiment 6	Macro			Positive			Negative		
	P	R	F1	P	R	F1	P	R	F1
NB (BoW)	0.64	0.62	0.60	0.70	0.42	0.52	0.58	0.82	0.68
NB (TF-IDF)	0.63	0.60	0.57	0.70	0.35	0.47	0.57	0.85	0.68
SVM (BoW)	0.71	0.71	0.71	0.73	0.67	0.70	0.69	0.75	0.72
SVM (TF-IDF)	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.69	0.70
KNN	0.66	0.53	0.40	0.51	**0.98**	0.68	0.81	0.07	0.12
RF	0.74	0.73	0.73	0.72	0.77	0.74	0.75	0.70	0.72
XGB (BoW)	0.75	0.75	0.75	0.73	0.79	0.76	0.77	0.71	0.74
XGB (WE)	0.80	0.79	0.79	0.78	0.82	0.80	0.81	0.77	0.79
SN	0.69	0.68	0.68	0.67	0.72	0.70	0.70	0.65	0.67
CNN	0.69	0.68	0.68	0.73	0.58	0.64	0.65	0.79	0.71
BiLSTM	0.69	0.68	0.68	0.66	0.76	0.70	0.71	0.60	0.65
TN	0.61	0.55	0.49	0.53	0.91	0.67	0.68	0.20	0.31
DistilBERT	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
RoBERTa	**0.86**	**0.86**	**0.86**	**0.86**	0.87	**0.87**	**0.87**	**0.86**	**0.86**
HerBERT	0.81	0.81	0.81	0.78	0.85	0.82	0.84	0.77	0.80
Polbert	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79

Experiment 7	Macro			Positive			Negative		
	P	R	F1	P	R	F1	P	R	F1
NB (BoW)	0.69	0.66	0.64	0.61	0.88	0.72	0.78	0.44	0.56
NB (TF-IDF)	0.80	0.66	0.61	0.59	**1.00**	0.74	**1.00**	0.31	0.48
SVM (BoW)	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
SVM (TF-IDF)	0.79	0.78	0.78	0.85	0.69	0.76	0.74	0.88	0.80
KNN	0.77	0.56	0.46	**1.00**	0.13	0.22	0.53	**1.00**	0.70
RF	0.88	0.88	0.87	0.93	0.81	0.87	0.83	0.94	0.88
XGB (BoW)	**0.92**	**0.91**	**0.91**	0.84	1.00	**0.91**	1.00	0.81	**0.90**
XGB (WE)	0.92	0.91	0.91	0.84	1.00	0.91	1.00	0.81	0.90
SN	0.25	0.50	0.33	0.50	1.00	0.67	0.00	0.00	0.00
CNN	0.90	0.88	0.87	0.80	1.00	0.89	1.00	0.75	0.86
BiLSTM	0.25	0.50	0.33	0.50	1.00	0.67	0.00	0.00	0.00
TN	0.50	0.50	0.38	0.50	0.06	0.11	0.50	0.94	0.65
DistilBERT	0.72	0.72	0.72	0.73	0.69	0.71	0.71	0.75	0.73
RoBERTa	0.88	0.84	0.84	0.76	1.00	0.86	1.00	0.69	0.81
HerBERT	0.77	0.56	0.46	1.00	0.13	0.22	0.53	1.00	0.70
Polbert	0.83	0.75	0.73	1.00	0.50	0.67	0.67	1.00	0.80

Experiment 8	Macro			Positive			Negative		
	P	R	F1	P	R	F1	P	R	F1
NB (BoW)	0.67	0.64	0.63	0.74	0.45	0.56	0.60	0.84	0.70
NB (TF-IDF)	0.68	0.64	0.61	0.77	0.39	0.52	0.59	**0.89**	0.71
SVM (BoW)	0.73	0.73	0.73	0.74	0.72	0.73	0.73	0.75	0.74
SVM (TF-IDF)	0.72	0.72	0.72	0.70	0.75	0.73	0.73	0.68	0.70
KNN	0.73	0.53	0.41	0.52	**1.00**	0.68	**0.95**	0.07	0.13
RF	0.75	0.75	0.75	0.74	0.78	0.76	0.77	0.72	0.74
XGB (BoW)	0.76	0.76	0.76	0.76	0.77	0.76	0.76	0.76	0.76
XGB (WE)	0.83	0.83	0.83	0.84	0.80	0.82	0.81	0.85	0.83
SN	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73
CNN	0.73	0.72	0.72	0.78	0.61	0.69	0.68	0.83	0.75
BiLSTM	0.73	0.72	0.72	0.75	0.66	0.71	0.70	0.78	0.74
TN	0.72	0.71	0.71	0.75	0.65	0.69	0.69	0.78	0.73
DistilBERT	0.84	0.84	0.84	0.84	0.83	0.83	0.83	0.84	0.84
RoBERTa	**0.88**	**0.88**	**0.88**	**0.86**	0.91	**0.88**	0.90	0.86	**0.88**
HerBERT	0.85	0.85	0.85	0.83	0.89	0.86	0.88	0.81	0.84
Polbert	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80

Experiment 9	Macro			Positive			Negative		
	P	R	F1	P	R	F1	P	R	F1
NB (BoW)	0.66	0.64	0.63	0.61	0.81	0.70	0.71	0.48	0.57
NB (TF-IDF)	0.67	0.64	0.63	0.61	0.83	0.70	0.72	0.46	0.56
SVM (BoW)	0.72	0.72	0.72	0.72	0.73	0.73	0.73	0.71	0.72
SVM (TF-IDF)	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73
KNN	0.75	0.51	0.35	**1.00**	0.02	0.03	0.50	**1.00**	0.67
RF	0.76	0.75	0.74	0.83	0.63	0.71	0.70	0.87	0.77
XGB (BoW)	0.72	0.72	0.72	0.72	0.73	0.73	0.73	0.71	0.72
XGB (WE)	**0.84**	**0.83**	**0.83**	0.79	**0.91**	**0.85**	**0.89**	0.76	**0.82**
SN	0.25	0.50	0.33	0.00	0.00	0.00	0.50	1.00	0.66
CNN	0.72	0.63	0.59	0.87	0.31	0.46	0.58	0.95	0.72
BiLSTM	0.25	0.50	0.33	0.00	0.00	0.00	0.50	1.00	0.66
TN	0.25	0.50	0.33	0.00	0.00	0.00	0.50	1.00	0.66
DistilBERT	0.65	0.65	0.65	0.63	0.68	0.66	0.66	0.61	0.63
RoBERTa	0.77	0.76	0.76	0.73	0.84	0.78	0.81	0.69	0.75
HerBERT	0.65	0.65	0.64	0.67	0.57	0.62	0.63	0.72	0.67
Polbert	0.66	0.65	0.65	0.63	0.75	0.68	0.69	0.56	0.62

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_15