# Deep Learning and Film History: Model Explanation Techniques in the Analysis of Temporality in Finnish Fiction Film Metadata

Filip Ginter[1], Harri Kiiskinen[2], Jenna Kanerva[1], Li-Hsin Chang[1] and Hannu Salmi[2]

[1]*TurkuNLP, Department of Computing, University of Turku, Finland*
[2] *Department of Cultural History, University of Turku, Finland*

### Abstract

We demonstrate the application of a deep-learning -based regressor, on a case study of predicting movie production year based on its plot summary. We show how the Integrated Gradients (IG) model explanation method can be used to attribute the predictions to individual input features and compare these to human-assigned attributions. Our purpose is to provide an insight into the application of modern NLP methods in the scope of a digital humanities research question, and test the model explanation techniques on a problem that is easy to understand, yet non-trivial for both humans and machine learning algorithms alike.

We find that the model clearly outperforms non-expert human annotators, being able to date the movies well within the correct decade on average. We also demonstrate that the model-assigned attributions agree with those assigned by humans, especially for correct predictions.

### Keywords

film history, deep learning, model explanation, text classification, NLP

## 1. Introduction

Deep learning (DL) methods have become the norm in Natural Language Processing (NLP), owing to their superior performance and versatility compared to the previous state of the art. A popular DL model architecture is the Transformer [1], which has been used to construct the Bidirectional Encoder Representations from Transformers (BERT) model [2], one of the most well-known variant of deep neural network models used very broadly in NLP.

Applications of DL methods in the digital humanities fields call not only for the better accuracy offered by these methods, but also for the ability to explain the predictions of these complex methods to a human researcher, in an interdisciplinary setting. These explanations are necessary to verify that the model bases its predictions on features meaningful in the context of the task, but can also be used to support further exploration of the data at hand and to

*The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022*

figint@utu.fi (F. Ginter); harri.kiiskinen@utu.fi (H. Kiiskinen); jmnybl@utu.fi (J. Kanerva); lhchan@utu.fi (L. Chang); hansalmi@utu.fi (H. Salmi)
https://fginter.github.io/ (F. Ginter)
0000-0002-5484-6103 (F. Ginter); 0000-0003-4187-5551 (H. Kiiskinen); 0000-0003-4580-5366 (J. Kanerva); 0000-0002-1301-2823 (L. Chang); 0000-0001-8607-6126 (H. Salmi)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3232/paper04.pdf

generate new hypotheses. Deep neural networks are commonly referred to as “black boxes” whose decisions are opaque to the user, and simple methods would be preferred as they are more directly interpretable.

In this paper, we demonstrate on a case study the application of the Integrated Gradients (IG) [3] model explanation method to a text classifier based on a BERT model. Our purpose is to provide an insight into the application of modern NLP methods in the scope of a digital humanities research question, and test the model explanation techniques on a problem that is easy to understand, yet non-trivial for both humans and machine learning algorithms alike.

Our domain of study is the Finnish National Filmography as preserved at the Finnish National Audiovisual Institute (KAVI)[1] and our data consists of the synopsis and other metadata for 1366 Finnish fiction films. This covers all fiction films of the National Filmography from the first film *Salaviinanpolttajat* (The Moonshiners, 1907) to ca. 2019, the current state of the filmography by June 2021. To gain a quantitative insight into the data, but equally importantly to demonstrate and benchmark the ability to explain DL model predictions, we use a rather straightforward proxy task: given the plot summary of the movie with years and peoples’ names masked, predict the production year of the movie. The task itself is obviously not trivial as the model is forced to learn the rather weak signals corresponding to the change of movie genres and typical plots over a full century. This, in turn, allows us to quantify these changes based on the explanations of the model predictions. Further, we are able to demonstrate the model explanation techniques applicable to complex DL models.

This study has practical relevance in the context of the *Movie Making Finland* -project. The dataset referred to above contains free-form textual data about the movies, but in contrast to the movies itself, this data is usually produced later. (See Section 2.2 below.) This causes source critical problems that may cause issues especially with machine learning tools: are the free-form texts more representative of their own times than of the things they are describing? This study will examine, if these textual data are actually able to deliver enough information about the original object of description for a machine learning algorithm to be able to date the movies even roughly.

In the following, we first describe the data and methods used, then introduce and analyze the model predictions, and finally carry out a human performance comparison study, comparing human- and model-assigned explanative features.

## 2. Data

### 2.1. Data source

The data is sourced from the Elonet database, including the Finnish National Filmography database, managed by the National Audiovisual Institute KAVI [4]. The database offers XML dumps of individual film metadata, and a simple crawler script was used to download these from the database user interface.

In order to facilitate further use of the metadata, the XML datasets were converted to semantic RDF data using an XSL template. This process was automated with a script. The data was

[1] https://elonet.finna.fi/

**Table 1**
Freetext fields in Elonet data (as of June 2021)

| Data | Amount |
|---|---:|
| Films | 1366 |
| Content Descriptions | 1237 |
| Synopses | 1166 |
| Commentaries | 1237 |

converted to RDF and stored in a project triple store to ease analysing movie metadata and to link analysis results to the movie information. The data is only described in parts relevant to the current study.

The Elonet data contains three fields with free-form texts describing the movies: content description, synopsis, and commentary. Since the Elonet database is a dynamic database, it does not provide a complete set of data for all movies. (See below for further discussion of this.) Table 1 shows a summary of the free-text data available.

There are as many content descriptions as commentaries, and somewhat fewer synopses, but in each case, the actual amount of data would be quite enough for the study. The contents of the fields differ considerably. The commentary places the movie into its wider context, it relates the movie to the contemporary social and cultural situation, and to the other works of the filmmakers; the synopsis, on the other hand, is a very short summary of the movie, one or two paragraphs, and consequently, offers relatively short texts. The content description is the most complete summary of the contents of the movie, describing the events and characters in detail, as well as the plot and its development.

Content descriptions, synopses and commentaries have mainly been written by the researchers of the National Audiovisual Institute. All Finnish fiction films before 1920, 27 films in total, have been lost. In these cases the content description derives from the time of the production: the text is either from the handout of the movie or from a description given to the censorship authorities. In all other cases the descriptions have been written by the researchers of KAVI. This field was chosen as the source for the experiments in this study, and the contents of this field are referred to as ”plot summary” in this paper. The actual dataset was obtained from the triple store with a simple SPARQL query. The data contained the movie id, the production year of the movie, and the contents of the content description field.

## 2.2. Assessment of the data quality

Usually movies are referred to using the year when they had their premiere, which is not necessarily the same as the year they were produced in. There are ten movies, where the production year in the database is different than the year of the premiere, and in all of these cases, the difference is one year.

For the plot summary data, the main question in relation to the aims of this study is, whether the analysis will pick stylistic features related to the time of the production of the text, or features of the actual plot summarized in the text. Is the movie recognized as old as the summary? The age of the description is not directly related to the movie in question. The writing of these content descriptions started when the Finnish National Filmography project was launched in the 1990s [5, p. 140]. The Filmography appeared in book form between 1996 and 2005 [6]. Therefore, most plot summaries should stylistically conform to what was considered a proper form of describing the films in the late 20th century; the data does not allow an accurate dating of individual descriptions, but in light of this information, it should be apparent, that any stylistic differences between plot summaries should, in fact, be results of how the proper description was produced in relation to the content of the movie, if any such stylistic differences can be found. The exception to this are the early, lost films (see above), for which there is way to create modern descriptions, and for which the content description relies on a contemporary text that has been deemed suitable for this purpose. If the analysis is based on stylistic matters, these films should stand out.

In addition to this, there is also the question of historical films as a genre. A considerable proportion of the film production concentrated on stories that described an era different from the time of the production. Finnish cinema has for example portrayed events of the 19th century or early 20th century. Thus, the content of the description includes features that refer to a completely different era than the time of the production. This brings forward the question if it still is possible to identify the year of the production.

What is the effect of the missing plot summaries? In Table 1 we see, that there are 129 movies without descriptions in the database. This missing data is the result of the backlog in documenting films. The film data is entered to the database as soon as possible, and completed as time and resources at the National Audiovisual Institute allow. This is evident when the missing data is analysed further: all missing descriptions are for films produced in 2013 or later; every film produced before that has a content description, *i.e.*, a plot summary.

## 3. Methods

Next, we introduce (a) the model we use to date the movies based on their plot summaries, and (b) the model explanation method we use to calculate the attributions of individual input features w.r.t. the prediction.

### 3.1. Text-to-year regression

We implement the regression of movie plot summaries to production year using the FinBERT [7] model. FinBERT is a monolingual pre-trained Finnish BERT model which has achieved many state-of-the-art results in Finnish NLP. It represents a class of modern NLP models based on deep neural networks and pre-trained on billions of tokens of raw language data. This pre-training step is one of the primary advantages of this class of models: unsupervised pre-training on a large quantity of text results in a model able to accurately encode input text. In the task-specific fine-tuning step, i.e. training the model for a particular task, this ability is transferred. In general, a BERT-based model would represent the default model choice in present-day NLP and can be reasonably expected to produce highly competitive results.

The model is illustrated in Figure 1. The output contextualized sub-word embedding vectors provided by the FinBERT model are averaged (mean pooling), and the regression is carried out by a straightforward linear projection layer from this embedding average. This follows the

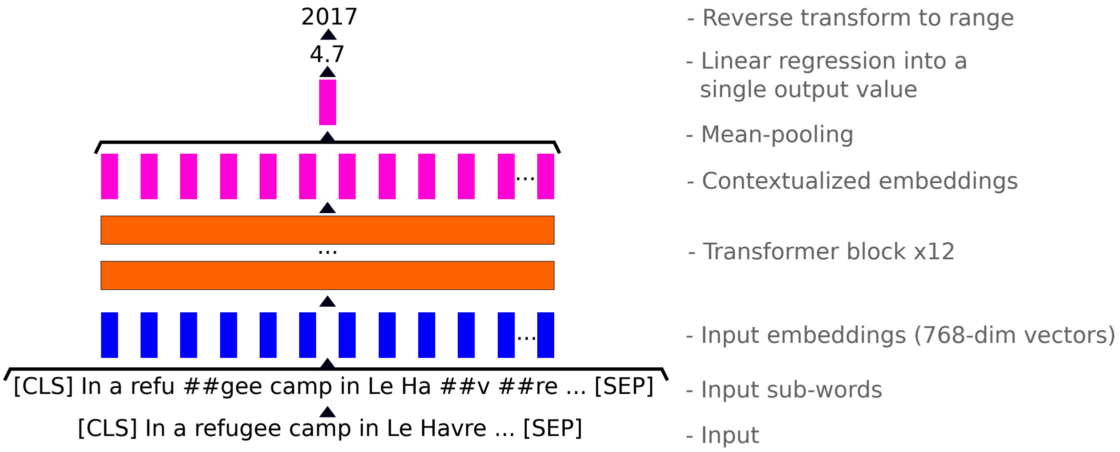

2017
4.7
[CLS] In a refu ##gee camp in Le Ha ##v ##re ... [SEP]
[CLS] In a refugee camp in Le Havre ... [SEP]
- Reverse transform to range
- Linear regression into a single output value
- Mean-pooling
- Contextualized embeddings
- Transformer block x12
- Input embeddings (768-dim vectors)
- Input sub-words
- Input

**Figure 1:** The BERT-based regression model used in this work. `[CLS]` and `[SEP]` are special marker tokens added to the input of the BERT model.

typical approach to applying the BERT model in different tasks: only a thin task-specific layer is applied on top of the BERT model. We experimented with several other alternatives, but these lead to worse accuracy on our development data, and we will not discuss these further in this paper. All weights of the model (including all BERT layers) are optimized during training. Based on initial experiments, the target of regression was not set directly to be the production year $y$, but rather its linear transformation $y' = \frac{y-1970}{10}$, which centers the data around its approximate mean of 1970 and squeezes its range, similar to z-transformation. Interestingly, and perhaps somewhat surprisingly, without initially centering and squeezing the data with this simple transformation, training the model was challenging and the performance of the model was poor. This was likely because the outputs of the randomly initialized linear regression layer were initially far from the target values and required a large number of training steps to reach the correct output range, while at the meantime the gradients passed to the BERT model caused its sub-optimal final performance. We have also tested the natural idea of first only optimizing the regression layer, and only after it reached the desired range continuing with optimizing the whole model. That approach mitigated most of the issues, but nevertheless resulted in clearly worse overall performance, and was not pursued further.

## 3.2. Feature attributions

There are numerous methods for establishing *feature attributions*, i.e. the assignment of importance to input features with respect to the prediction made by the model. In this work, we apply Integrated Gradients (IG) [3], as a popular, representative example of such methods, specifically targeting differentiable models such as deep neural networks. In short, the IG method defines the attribution of a feature as the integral of the gradient of the model output w.r.t. the given feature along the path from a “blank” reference input to the actual input. In practice, it is implemented by interpolating the model input between the reference input and the actual input in $N$ steps (here we set $N = 50$) thus evaluating the model $N$ times. In image processing, the reference input would be e.g. an empty image. In this work, we use the sequence `[CLS] [PAD]`

[PAD] ... [PAD] [SEP] to represent a "blank" input of a BERT-based classification model: [CLS] and [SEP] are the special separation tokens in BERT input, and [PAD] is the padding token. This reference sequence has same length as the actual input and the interpolation is carried out on the input token embedding vectors. To get an attribution value for each input token, we simply sum up the attributions across the 768 dimensions of each input embedding. A positive attribution value signals contribution *towards* the prediction made by the model, while a negative attribution value signals contribution *against* the prediction made by the model.

Finally, the BERT model uses sub-word tokenization, which splits rare words into sub-words, so as to maintain a fixed-length vocabulary (see illustration in Figure 1). An input of $N$ words (in the usual sense) is therefore presented to the model as a sequence of $M$ sub-words, where typically $M > N$. In order to obtain word-level attributions, understandable to the human reader, we set the attribution of a word to be the attribution of that of its sub-words which has the highest absolute value. Thus, for instance, if an input word is divided into three sub-words with attributions of $[-0.4, 0.1, 0.21]$, the overall attribution of the word will be $-0.4$.

### 3.3. Data pre-processing

Other than sub-word tokenization, the BERT model does not require any particular input pre-processing. A somewhat unfortunate property of the BERT model is its maximum input length of 512 sub-words for any given input sequence. In this work, we trim the plot summaries to fit the maximum sequence length of the model. On average, this preserves 79% of the content description length and we therefore did not see any need for a more complicated solution.

So as to avoid accidentally revealing features in the inputs in the form of production years and/or character names for both machine-learning and human experiments, we replace all digits in the data with the numeral 0 and all people names with the BERT special token [MASK]. The names are recognized using the TurkuNLP Finnish NER system [8], a state-of-the-art system for Finnish named entity recognition (NER).

### 3.4. Baselines

When reporting the results, we consider several baselines. As trivial baselines, we use the constant prediction baseline that predicts the mean production year in the data for every item and the random baseline that predicts a random year from the year range in the data.

We also test the linear support vector regressor (SVR) as a representative of linear methods often applied in text classification tasks for their simplicity, light computational demands, and straightforward explainability through the linear feature coefficients learned by the model.

### 3.5. Experimental setup and parameters

Throughout the evaluation, we use a single randomized split to 80% (990 examples) training data, 10% (123 examples) development data, and 10% (124 examples) test data. All parameter selection is carried out on the development data. The human baseline experiment described below is carried out on the first 20 movies in the test data. The optimized loss is mean square error of the regressed value. The final parameter settings were: batch size 30, gradient accumulation over 3 batches, learning rate 5e-5, maximum number of training steps of 1500, warm-up of 150 steps, and early-stopping with patience of 3 and evaluation every 30 steps. For the best-parameters model, early-stopping triggered after 540 training steps, corresponding to nearly 50 epochs of training. All other parameters were in their default value as set in the torch-backend of the Hugging Face transformers library version 4.15.0 [9]. The experiments were carried out using GPU-accelerators in the CSC super-computer centre accessible to all of Finnish academia. A single training run took approximately 30 minutes and attribution calculation 1.2 seconds per example, all on a high-end GPU accelerator (AMD MI100). FinBERT is of the *base* BERT variant, with 12 Transformer layers and embedding width of 768 dimensions.

The linear SVR baseline parameters are grid-searched on the development data. The best combination is C value of 10, TF-IDF weighted 2–5-character n-gram tokenization representing word boundaries, and 300,000 unique features. The baseline is implemented using the *scikit-learn* library version 1.0.2 [10].

We use two primary metrics to evaluate the model performance: The *mean absolute error* (MAE) indicates how many years the prediction was mistaken on average regardless of the direction of the misprediction, and is calculated as the mean of the absolute values of the prediction errors. The *mean error* (ME) then captures possible biases in predictions, negative values indicating biases towards assigning older years while positive values indicate biases towards recent years.

# 4. Results

## 4.1. Model performance

The predictions of the BERT model on the blind test set are shown in Figure 2. The mean error (ME) is 0.012 years, which can be interpreted as there not being an overall bias towards over- or under-estimating the age of the movie. The mean absolute error (MAE) is 7.43 years, or in other words the model can place the movie well within the correct decade, based on its plot.

Figure 2 shows that both the data and the errors are evenly distributed without any particular clear biases, except for the expectable fact that errors in predictions for very old movies are positive, and errors in predictions for very new movies are negative. That is merely a consequence of the fact that the model has learned the overall range in the data, and naturally does not predict outside this range. For comparison, the mean absolute error of the trivial baseline which predicts a constant value equal to the mean of the test set is 25.2 years and the linear SVR achieves MAE of 10.24. Even though the BERT model reduces the linear model's error by a full 27%, the result of the linear baseline is nevertheless very good and very clearly above the mean baseline.

In our test set of 124 films, there are 35 films with a deviation of 10 years or more. In 21 cases the model predicted the film to be older than it actually is, while in 14 cases the predicted year was not old enough. It seems that the model had particular difficulties in judging the production year for the films from the 1920s. There were seven silents from 1922–1929 which were estimated to be from 1934–1955. Most of these films were rural melodramas and represented the kind of production trend that was also typical of the 1930s, 1940s and 1950s in Finland. This might be a background for the misinterpretation since the content descriptions of these films did not include such features that could be regarded characteristic of the 1920s only. In six of these

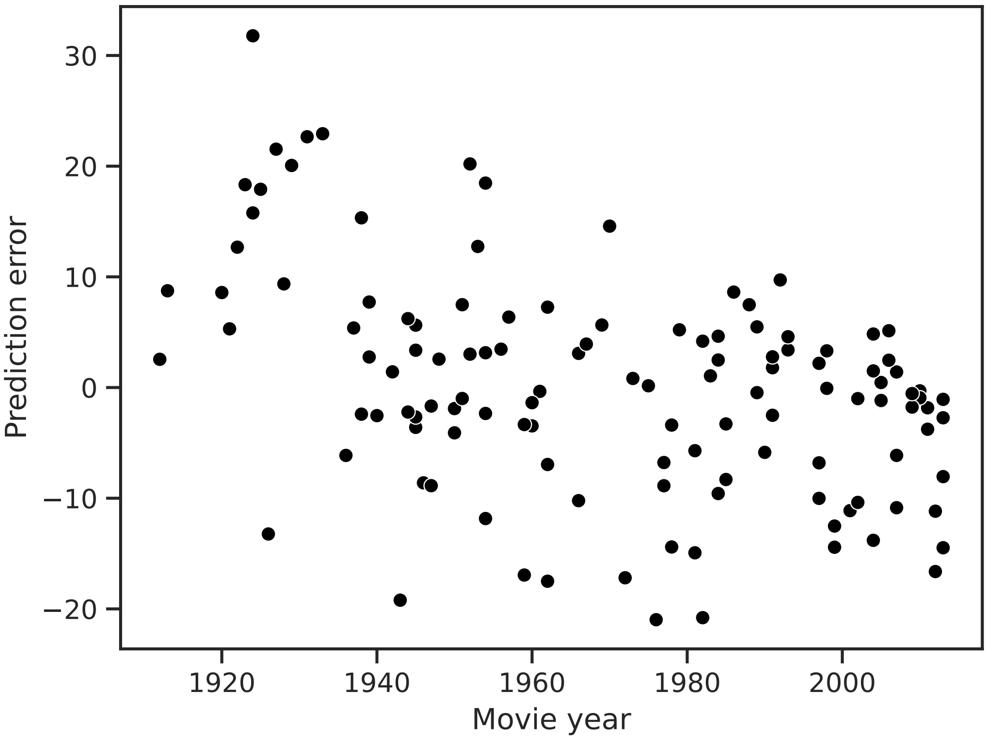


**Figure 2:** Prediction errors of the model, plotted against the actual year of each movie in the test set.

films, the gap is between 12.68–21.54. The seventh film *Polyteekkarifilmi* (1924) is a special case, with the widest gap in the whole test set, 31.78 years. The film itself was exceptional in its time, an amateur film, with many advertisements and other commercial inserts. The story was set both in the 1920s Helsinki and on the Olympus Mountain, from where ancient gods decide to depart for Finland.

In 21 films the gap is ten year or more backwards in time. This group is more heterogeneous, comprising films from nine different decades. Ten of these films are however from the 1990s and 2000s, and in all these cases the model has predicted the film to be considerably older. There is no one reason behind this, but it seems that in this group there are many historical films, where the story has been set into the past, for example *Lapin kullan kimallus* (1999) which takes place in the late 19th century. Obviously, in the content descriptions of historical films the text includes elements that differ from the discourse of the production year.

## 4.2. Comparison to human performance

In order to better interpret the model performance in context of the task, we compare the regression model to an average human performance. To this end, we randomly sample 20 movies

**Table 2**
Mean Error (ME) and Mean Absolute Error (MAE) evaluated on the 20 movies randomly sampled from the test set.

| | ME | MAE |
|---|---|---|
| Regression | -0.2 | 8.4 |
| Baseline (constant) | 0.0 | 26.8 |
| Baseline (random) | -21.5 | 37.3 |
| Baseline (linear) | -0.1 | 10.2 |
| Annotator 1 | 1.9 | 23.6 |
| Annotator 2 | 11.4 | 14.8 |
| Annotator 3 | 3.0 | 16.0 |
| Annotator 4 | 11.8 | 27.0 |

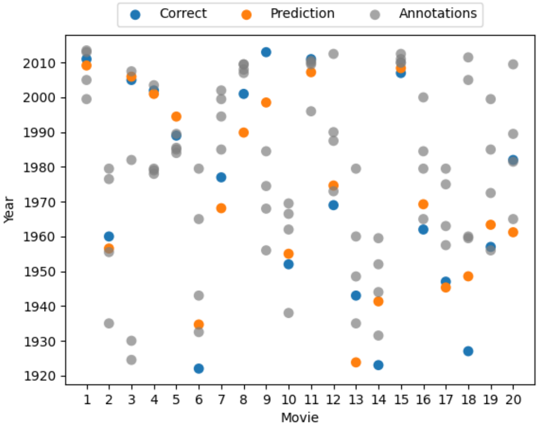


**Figure 3**
The correct production year, the year predicted by the regression model, as well as the human made annotations visualized for each 20 movies in the random sample.

from the test set, and ask four human annotators independently to estimate the production year for each movie from the given plot summary. We choose non-expert annotators rather than movie historians to avoid correct judgments based on deep knowledge of the Finnish movie history. Similarly to the regression model input, all numbers and person names appearing in the text are masked in order to avoid basing the estimate on the reported years and recognizable names in case such happen to appear in the text. In addition to the production year, the annotators are asked to select five words from each plot description that in their opinion contributed most to their final decision, and thus serving as keywords for explaining the human annotations.

The evaluation of the human performance compared to the regression model as well as the baselines (mean, random, and linear) is shown in Table 2 in terms of mean error (ME) and mean absolute error (MAE). The results show all annotators having at least a slight bias towards recent years, while the model is able to predict the correct distribution more precisely. In terms of mean absolute error, the regression model is clearly better than any of the human annotators, the model on average having a mistake of approx. 8 years while the human annotators have average mistakes between 15–27 years.

In Figure 3 we plot the correct production year, the production year predicted by the regression model, and the four human annotations for all 20 movies in the manual sample. While some of the movies receive both the prediction and all annotations close to the correct production year (especially movies numbered 1, 5, 11, and 15), for most of the movies the human estimated production years vary greatly, not showing a clear consensus towards any decade. Especially interesting are movies numbered 3 and 18, where the human estimation varied between 1925 and 2007 (correct production year being 2005) and between 1960 and 2011 (correct production year being 1927) respectively.

When considering these results, it is important to note that both the regression model and the human annotation have an advantage over the other. While the regression model was exposed to the 990 movie plot summaries from the training data and is therefore able to learn the distribution and common characteristics from these, the human annotators have not studied the training data. However, the annotators naturally possess world knowledge and they may also recall the test movies from other context and thus recognize the movie based on its description. The four annotators were able to recognize 1, 5, 3, and 0 movies respectively based on the plot summary, however, the production year was estimated correctly only once among these recognized movies and the largest error marginal among these was as much as +22 years. However, in most of these cases the recognition likely helped to place the movie near the correct decade, and without these examples the human performance would have been even worse.

## 4.3. Keyword analysis

One of the key questions we ask in this paper is to what extent the predictions of the deep neural network -based model can be explained in terms of attributions to features in the input text. To this end, we next compare the keywords obtained from the regression model using the IG method to those manually selected by the annotators. While the annotators were instructed to select five keywords for each movie, the IG method produces a list of input words weighted by their estimated relevance, and therefore we are able to select any number of keywords for the model. Here, we select the top 10 keywords for each movie, skipping tokens composed of pure punctuation characters.

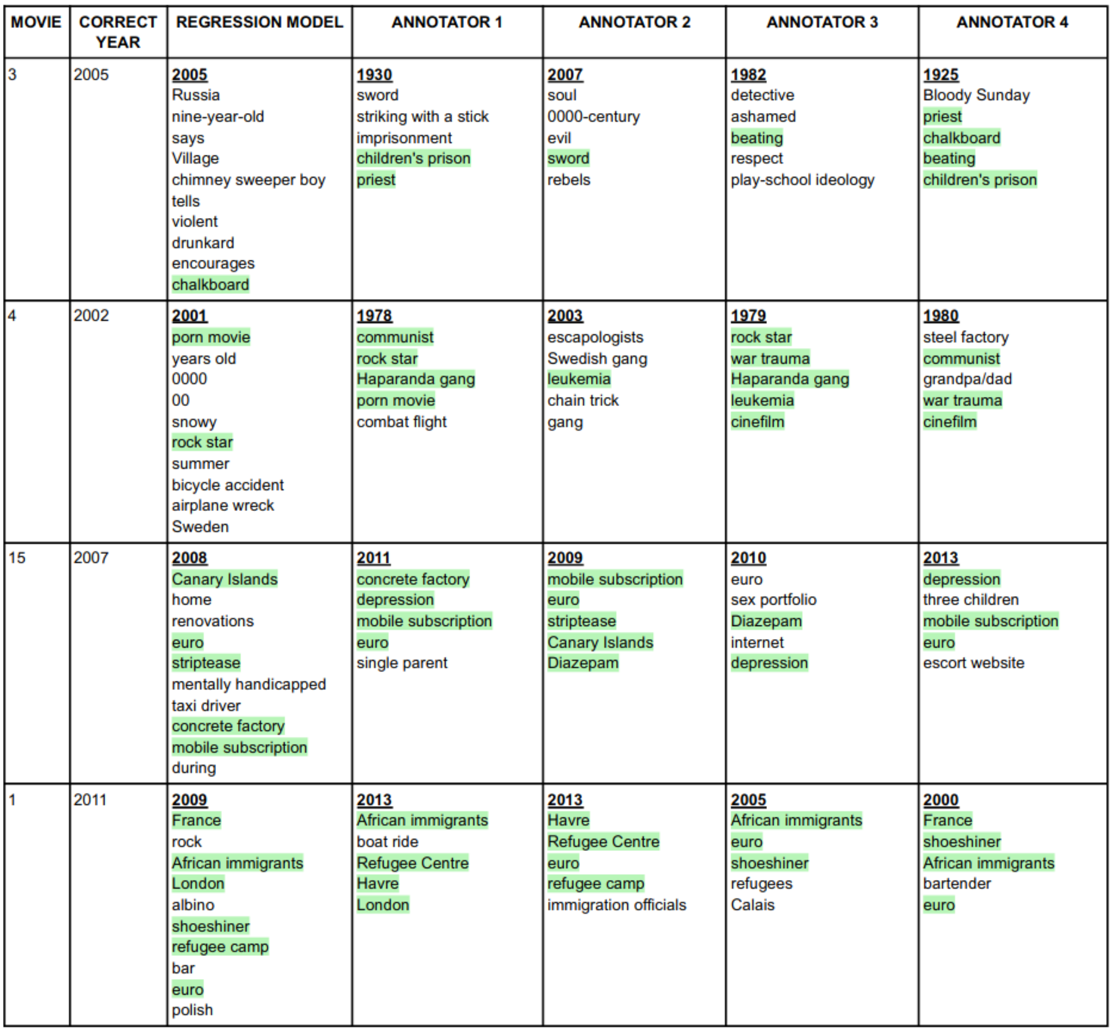
In Table 3 we visualized the keywords for the four movies from the manually annotated sample that had the smallest absolute prediction error of the regression model. All keywords selected from the same movie by at least two different annotators or an annotator and the regression model are highlighted with green, showing a substantial overlap in the extracted keywords. In the two first rows in the table (movies numbered 3 and 4) the overlap of model and human extracted keywords is quite low, likely due to most of the human predicted production years being quite underestimated predicting productions years much older than the actual year, therefore decreasing the quality of the selected keywords as well. However, in the two last rows (movies numbered 15 and 1) human estimates are closer to the actual production years, and there is a substantial overlap between the model and human extracted keywords.

When numerically comparing over the manually annotated sample of 20 movies, considering top-10 keywords as estimated by the regression model, 75% of the movies have at least one keyword in common with those extracted by the annotators, and when increasing to top-20 keywords at least one overlapping keyword is found from 85% of the movies.

Overall, both the regression model as well as the annotators are capturing topical concepts easily pointable to specific time. Such concepts include terms relating to technical development (cell phone, microcar), currency (euro, Finnish markka), or old-fashioned titles and professions not as widely used in the modern society (health-sister (literal translation of an old Finnish term for a nurse), lensmann).

Finally, as a "sanity check" experiment, we modified the dataset so as to include the sentence *"The movie was filmed in NNNN."* at a randomly selected position in the plot summary (candidate positions are on sentence boundaries so as to preserve elementary text flow). Here NNNN is the correct year of production for each movie, revealing the correct output to the regressor. As

**Table 3**
The keywords extracted from the regression model as well as selected by the annotators for the four movies from the manually annotated sample that had the smallest absolute prediction error of the regression model. All keywords appearing at least twice in the same movie are highlighted with green, and all keywords are translated into English.



| MOVIE | CORRECT YEAR | REGRESSION MODEL | ANNOTATOR 1 | ANNOTATOR 2 | ANNOTATOR 3 | ANNOTATOR 4 |
|---|---|---|---|---|---|---|
| 3 | 2005 | **2005**<br>Russia<br>nine-year-old<br>says<br>Village<br>chimney sweeper boy<br>tells<br>violent<br>drunkard<br>encourages<br>chalkboard | **1930**<br>sword<br>striking with a stick<br>imprisonment<br>children's prison<br>priest | **2007**<br>soul<br>0000-century<br>evil<br>sword<br>rebels | **1982**<br>detective<br>ashamed<br>beating<br>respect<br>play-school ideology | **1925**<br>Bloody Sunday<br>priest<br>chalkboard<br>beating<br>children's prison |
| 4 | 2002 | **2001**<br>porn movie<br>years old<br>0000<br>00<br>snowy<br>rock star<br>summer<br>bicycle accident<br>airplane wreck<br>Sweden | **1978**<br>communist<br>rock star<br>Haparanda gang<br>porn movie<br>combat flight | **2003**<br>escapologists<br>Swedish gang<br>leukemia<br>chain trick<br>gang | **1979**<br>rock star<br>war trauma<br>Haparanda gang<br>leukemia<br>cinefilm | **1980**<br>steel factory<br>communist<br>grandpa/dad<br>war trauma<br>cinefilm |
| 15 | 2007 | **2008**<br>Canary Islands<br>home<br>renovations<br>euro<br>striptease<br>mentally handicapped<br>taxi driver<br>concrete factory<br>mobile subscription<br>during | **2011**<br>concrete factory<br>depression<br>mobile subscription<br>euro<br>single parent | **2009**<br>mobile subscription<br>euro<br>striptease<br>Canary Islands<br>Diazepam | **2010**<br>euro<br>sex portfolio<br>Diazepam<br>internet<br>depression | **2013**<br>depression<br>three children<br>mobile subscription<br>euro<br>escort website |
| 1 | 2011 | **2009**<br>France<br>rock<br>African immigrants<br>London<br>albino<br>shoeshiner<br>refugee camp<br>bar<br>euro<br>polish | **2013**<br>African immigrants<br>boat ride<br>Refugee Centre<br>Havre<br>London | **2013**<br>Havre<br>Refugee Centre<br>euro<br>refugee camp<br>immigration officials | **2005**<br>African immigrants<br>euro<br>shoeshiner<br>refugees<br>Calais | **2000**<br>France<br>shoeshiner<br>African immigrants<br>bartender<br>euro |

expected, this results in near-perfect predictions and the token with the year is flagged as the single top-most important feature in 60% of all predictions.

## 5. Conclusions and future work

In this work, we set out to demonstrate the applicability and explainability of a modern, DL-based NLP approach to text-to-value regression on a case study in the digital humanities domain.

Firstly, we find that a modern BERT-based regressor is capable of highly accurate predictions that are substantially more accurate than (non-expert) humans. More importantly for the digital humanities domain, though, the model predictions can be successfully traced to individual input text tokens so as to give insight into the predictions. Further, we show that the individual tokens attributing to the model's predictions agree with manually selected keywords, especially for predictions where both humans and the model are close to correct.

Secondly, we gained interesting insights on the task itself, gaining understanding on the expected predictability of movie plots in time, finding out that, even with names and numbers masked, a surprisingly accurate prediction can be made on average, better than we intuitively expected prior to the study. Further, we could observe that the model based its predictions on the topical concepts in the descriptions, as would be intuitively expected.

In future work, a broader analysis will be carried out using predictions on all available data and we will attempt to aggregate the keyword explanations in time and compare it to topic-in-timeline types of models. Further, we will explore other model explanation methods in addition to the Integrated Gradients.

Both code and data are available at https://github.com/MoMaF/momaf_regressor.

## Acknowledgements

This work has been supported by the research consortium *Movie Making Finland: Finnish fiction films as audiovisual big data, 1907–2017* (MoMaF), funded by the Academy of Finland (329266). Film data and metadata were provided by the National Audiovisual Institute in Finland. Computational resources were provided by the CSC – IT Center for Science.

## References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[3] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, 2017, p. 3319–3328.

[4] Kansallinen Audiovisuaalinen Instituutti, Suomen kansallisfilmografia, 2020–.

[5] M. Kuutti, From film to stream – KAVI's Elonet, ten years in the making, Journal of Film Preservation (2020) 137–144.

[6] K. Uusitalo, S. Toiviainen (Eds.), Suomen kansallisfilmografia, Edita ja Suomen elokuva-arkisto, Helsinki, 1996–2005.

[7] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, S. Pyysalo, Multilingual is not enough: BERT for Finnish, arXiv preprint arXiv:1912.07076 (2019).

[8] J. Luoma, M. Oinonen, M. Pyykönen, V. Laippala, S. Pyysalo, A broad-coverage corpus for Finnish named entity recognition, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 4615–4624. URL: https://aclanthology.org/2020.lrec-1.567.

[9] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, G. Louppe, Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2012).