

Extractive Text Summarization Using Generalized Additive Models with Interactions for Sentence Selection

Vinícius Camargo da Silva^a, João Paulo Papa^b and Kelton Augusto Pontara da Costa^c

São Paulo State University - UNESP, Bauru, Brazil

Keywords: NLP, Text Summarization, Interpretable Learning.

Abstract: Automatic Text Summarization (ATS) is becoming relevant with the growth of textual data; however, with the popularization of public large-scale datasets, some recent machine learning approaches have focused on dense models and architectures that, despite producing notable results, usually turn out in models difficult to interpret. Given the challenge behind interpretable learning-based text summarization and the importance it may have for evolving the current state of the ATS field, this work studies the application of two modern Generalized Additive Models with interactions, namely Explainable Boosting Machine and GAMI-Net, to the extractive summarization problem based on linguistic features and binary classification.

1 INTRODUCTION


Nowadays, computer systems have assumed important role in providing useful information inside the increasingly amount of data generated daily. In this context, Natural Language Processing (NLP) have gained space and applicability as a result of considerable amounts of text distributed across news portals, social media and various sources. With the growth of textual data, finding exact information may be a difficult task (El-Kassas et al., 2020). As such, Automatic Text Summarization (ATS) is becoming relevant (El-Kassas et al., 2020), as it fosters automatic strategies for building textual comprehensible summaries, ideally capable of preserving the original content and meaning (Moratanch and Chitrakala, 2017) while distilling the most important information considering the user involved (Maybury, 1995).


Some of the previous ATS approaches were simpler and had certain transparency, for example, using linear equations or fuzzy rules with statistical or linguistic features to map the importance of document sentences to produce extractive summaries (Mutlu et al., 2019; Afsharizadeh et al., 2018). In general, the simpler and the clearer the predictors used, the more directly it is possible to observe how the summary sentences are being chosen in case of extractive summarizers.


However, not all available approaches are transparent when considering model's decisions, especially in the machine learning context. With the popularization of public large-scale datasets, various deep learning approaches have been explored during the last years, competing for the state-of-the-art on such data. Despite their notable ability to elaborate summaries, as a problem inherited from dense architectures and language embeddings opacities (Danilevsky et al., 2020), in practice, such models usually yield shallow or obscure interpretations about their true behaviour and decisions.

Machine learning applications have been successful in different areas based on their statistical accuracy, but often lack clarity when explaining how decisions are actually being made. The so-called XAI (or Explainable Artificial Intelligence) field focuses on the study and elaboration of more explainable AI methodologies (Došilović et al., 2018; Samek and Müller, 2019; Arrieta et al., 2020). Depending on the application, explainability could impact in different aspects of a model, as its lack of interpretability can undermine its trustworthiness towards users or even hide potential improvements (Danilevsky et al., 2020; Arrieta et al., 2020).

Within the context of ATS, interpretable modelling relates to giving transparency to the model's summarization process, which could contribute to a better sense of the limitations and capabilities of the model, help the investigation of why the model makes mistakes or even assist in gaining insights about the problem itself. Such information could be useful, for

^a  <https://orcid.org/0000-0002-5327-0747>

^b  <https://orcid.org/0000-0002-6494-7514>

^c  <https://orcid.org/0000-0001-5458-3908>

example, for evolving approaches and for clarifying what the model actually satisfy in contrast to user's expectations.

With the interest in moving towards interpretable ATS learning models, this work aims to study the application of Generalized Additive Models with Interactions (GAMI) to the extractive summarization problem. More specifically, this work investigates training Explainable Boosting Machine (EBM) and GAMI-Net models in a binary classification fashion to later inferring the relevancy of sentences in the documents of interest.

To the best of our knowledge, this is the first application of EBMs or GAMI-Nets as the decision algorithm involved on extractive summarization. EBM and GAMI-Net models are built, respectively, on ensemble of trees and neural networks, whose main attempt is to balance intelligibility and accuracy in supervised problems combining main effects and pairwise interactions additively. The idea is to benefit from the additive formulation to make the behaviour and contributions explicit considering explanatory features and outputs, allowing intelligibility along the process. This work evaluates these models for news summarization (CNN/Dailymail) and long document summarization of scientific papers (PubMed), comparing to other machine learning algorithms and recent approaches.

2 BACKGROUND

Automatic Text Summarization is an NLP task that grows in importance with the expansion of data in textual form and the interest in exploring it efficiently, since assisting users understanding over documents could save time and effort (El-Kassas et al., 2020).

As pointed by (Luhn, 1958), the summarization process may require familiarity with the subject, which could culminate in qualified human resources dedicated to facilitating access to information. Therefore, the importance of ATS relies on the potential of reducing human efforts while accelerating reading time over text sources (Moratanch and Chitrakala, 2017) provided by automatic summarization.

A common way to distinguish text summarization approaches is between extractive and abstractive modelling. On extractive summarization, the model selects parts of the source text to compose the projected summary (Nenkova and McKeown, 2011), whereas on abstractive summarization the model may reuse parts of the source, but new terms and sentences are expected to appear (Nenkova and McKeown, 2011).

Traditionally, extractive techniques have the advantage of not suffering from grammatical or semantical issues in the summary (Nallapati et al., 2017), as usually the approaches are based on selecting entire sentences from the source text, leading to faster and simpler methods than abstractive techniques (El-Kassas et al., 2020). One way to give general description of the extractive summarization process can be done through three important steps (Nenkova and McKeown, 2012):

1. Creating an intermediate representation of the input;
2. Scoring sentences based on this representation; and
3. Selecting sentences to compose the summary.

Previous work such as the scoring approach presented by (Afsharizadeh et al., 2018) and the fuzzy systems addressed by (Mutlu et al., 2019) can be seen as examples of extractive summarization techniques with some level of interpretability for human intuition. The authors in (Afsharizadeh et al., 2018) relied on experimentally setting feature weights to linear feature functions and ranking sentences according to the obtained scores, while (Mutlu et al., 2019) presented fuzzy systems that distinguish summary-worthy and summary-unworthy sentences with the help of a fully data-driven rule generation scheme, providing an interesting alternative to manually generating rule sets.

Supervised machine learning algorithms emerged as an option for modelling extractive summarization that offers statistical performance at the price of requiring a dataset of considerable size (Wong et al., 2008; El-Kassas et al., 2020). Considering the interest on fully explainable models (Arrieta et al., 2020; Danilevsky et al., 2020; Samek et al., 2020), bringing explainability to ATS is becoming a necessity (Sarkhel et al., 2020). Being able to produce summaries through interpretable approaches may be an important step for machine learning-based models; however, literature focused on this topic is still scarce.

Recently, the authors in (Ghodratnama et al., 2020) combined supervised and unsupervised learning into a model called ExDoS that learns features weights as part of an extractive summarization approach. Such weights are discussed by the authors as a form to indicate the learned importances of the features, which may help the interpretation of how the model is deciding to select important sentences and avoid unimportant ones when elaborating summaries.

With a similar intention, in this work, we present an attempt to build interpretable extractive summarization models by taking advantage of Generalized Additive Models with Interactions, relying on their

transparency concerning features and outputs, to give intelligibility to the sentence selection process.

2.1 Generalized Additive Models with Pairwise Interactions

Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1987) are statistical approaches that have gained interest for their ability to be intelligible, appealing to their additive formulation and human intuition to provide interpretability in supervised problems.

These models have the form of Equation 1, working as middle ground between linear models (e.g., linear regression) and full-complexity models (e.g., ensembles of trees) (Lou et al., 2012), where the motivation could be, for example, obtaining predictors more accurate than the former while staying more intelligible than the latter, fitting individual features with non-linear functions that are combined additively (Lou et al., 2012):

$$g(y) = \sum_{i \in N} f_i(x_i), \quad (1)$$

where x_i is the i -th feature given a set N of features, f is called *shape function*, and g is called *link function*.

On additive models, the contribution of each feature map towards final decisions can be seen more clearly than in dense ones. This property allows GAM's learned shape functions to be visualized and their outputs to be investigated for individual or group of predictions, providing interpretability over the model and its decisions. Moreover, as GAMs can assume non-linear behaviour through their shape functions, they may fit a wider variety of problems when compared to linear models.

Recently, some works were focused on effectively building more powerful GAMs with the addition of modern machine learning strategies and supplementing the original univariate shape functions with a usual restricted set K of pairwise interactions (bivariate shape functions), increasing the accuracy of the final model while maintaining some intelligibility, resulting in the following models:

$$g(y) = \sum_{i \in N} f_i(x_i) + \sum_{(i,j) \in K} f'_{ij}(x_i, x_j), \quad (2)$$

where f' is a pairwise interaction.

In this work, two representatives of those algorithms are applied to the extractive summarization problem, i.e., EBM and GAMI-Net.

2.1.1 Explainable Boosting Machine

Explainable Boosting Machines (also known as Generalized Additive Model plus Interactions (GA²M) (Lou et al., 2013)) are modern tree-based GAMs with pairwise interactions based on ensembles, which achieved accuracy comparable to full-complexity models while keeping interpretability similar to former GAMs (Lou et al., 2012; Lou et al., 2013; Nori et al., 2019).

Given a supervised problem, tree-based GAMs can be learned upon residuals using gradient boosting, cycling through the features and improving shape functions iteratively (Lou et al., 2012). Using this algorithm with bagged trees have led to even better results on different regression and binary classification datasets (Lou et al., 2012).

Another improvement was addressed by (Lou et al., 2013), encompassing the inclusion of pairwise interactions into tree-based GAMs. Considering computational cost, the authors propose an approach based on firstly building an additive model with only univariate shape functions, and then ranking and selecting a fixed number of pairwise interactions that are fit on the residuals. The model that wraps those improvements was later called *Explainable Boosting Machine* by (Nori et al., 2019).

2.1.2 GAMI-Net

GAMI-Net is an interpretable neural network based on GAMs with structured interactions (Yang et al., 2021). The architecture consists in additive subnetworks with multiple hidden layers, each of which capturing a different shape function of Equation 2. The idea is to produce a model that keeps interpretability aspects of GAMs while relying on the power of deep neural networks to model non-linear behavior of shape functions.

The authors in (Yang et al., 2021) proposed an adaptive training algorithm using mini-batch gradient descent that fits main effects and pairwise interactions in separated stages. Firstly, the algorithm train the main subnetworks, pruning the trivial ones according to their contributions. Secondly, the algorithm selects a fixed number of pairwise interactions using the ranking procedure proposed by (Lou et al., 2013) and fit their respective subnetworks on the residuals, pruning trivial ones as in the first stage. Lastly, in a third and final stage, all the network parameters are fine-tuned together.

Moreover, GAMI-Net is designed to preserve sparsity, heredity and marginal clarity considering the main effects and pairwise interactions by keeping only non-trivial shape functions, including pairwise

interactions only if at least one of their parent main effects are kept and putting a regularization factor that enforces main effects and pairwise interactions to be more identifiable, which is intended to contribute to the model's interpretability (Yang et al., 2021).

3 PROPOSED APPROACH

Let $D = \{s_0, \dots, s_n\}$ be a document composed of a sequence of sentences s_i . The goal of our extractive summarization process is to obtain a sequence S of the most relevant sentences in D , where S is limited in size to be shorter than D .

We train EBM and GAMI-Net models using a set of six features from the literature to be able to rank and select sentences to compose S . The approach consists in training a model to decide which sentences are summary-worthy and then using such model to select the most important sentences in the input document to compose summaries. The simplicity of the explored features should contribute to both model efficiency and intelligibility. The approach is divided in preprocessing, feature extraction and sentence scoring and selection, as follows.

3.1 Preprocessing

The preprocessing step is responsible for turning raw documents into sequences of sentences, capturing useful information for future feature extraction. The process starts by segmenting raw documents into text sentences, which are split in tokens. Then, the process follows by removing punctuation and stopwords, word tagging and stemming. After preprocessing, sentences correspond to lists of their respective words that are forwarded to the feature extraction step.

We perform most of the NLP preprocessing with the Python library spaCy (Montani et al., 2021), with the exception of stemming step which is done using the NLTK SnowballStemmer (Bird et al., 2009).

3.2 Feature Extraction

After sentences are preprocessed, six different features are extracted from sentences to become inputs vectors $x = \{x_1, x_2, \dots, x_6\}$ that are used to train and predict. The feature computations are formulated as described below:

3.2.1 TF-ISF

TF-ISF is a variant of the TF-IDF method applied at sentence level for text summarization (Oliveira et al.,

2016; Mutlu et al., 2019). The idea is to compute a score to each sentence based on term importance and descriptiveness inside the document (Oliveira et al., 2016), which are measured by term frequency (TF) and inverse sentence frequency (ISF) of the terms. We use bigrams TF-ISF, so each sentence s_i of a document receives a salience score (Equation 4) based on its term bigrams b :

$$w(s_i) = \sum_{j=1}^{J_i} \left[F(b_j) \times \log \left(\frac{n}{nb_j} \right) \right], \quad (3)$$

$$x_1(s_i) = \frac{w(s_i)}{\max(w(s_i))}. \quad (4)$$

where $F(b)$ is the frequency of b in the document, n is the number of sentences in the document, nb is the number of sentences of the document in which b occurs and J_i is the number of bigrams s_i .

3.2.2 Position

Depending on a document type, how early or how late sentences appear may give important information about their relevancy (Ferreira et al., 2013; Oliveira et al., 2016). The position feature (Equation 5) represents the sentence position inside the document, where p_i is the position of sentence s_i :

$$x_2(s_i) = \frac{p_i}{n}. \quad (5)$$

3.2.3 Length

The length feature (Equation 6) is calculated based on the length of sentence s_i in terms of the maximum sentence length. The length feature allows the model to learn the relationship between sentence length and relevancy (Oliveira et al., 2016; Mutlu et al., 2019):

$$x_3(s_i) = \frac{\text{number of terms in sentence } s_i}{\max(\text{number of terms in a sentence})}. \quad (6)$$

3.2.4 Proper Nouns and Numerical

The individual ratio of proper noun and numerical terms in the sentence s_i may indicate the presence of relevant information (Oliveira et al., 2016). We calculate these features as follows:

$$x_4(s_i) = \frac{\text{number of proper nouns in } s_i}{\text{number of terms in } s_i} \quad \text{and} \quad (7)$$

$$x_5(s_i) = \frac{\text{number of numerical terms in } s_i}{\text{number of terms in } s_i}. \quad (8)$$

3.2.5 Sentence-Sentence Similarity

The sentence-sentence similarity denotes how close a sentence is to other sentences in the document (Mutlu et al., 2019). We calculate this feature using cosine similarity c as in Equation 9:

$$x_6(s_i) = \frac{\sum_{j=1}^n c(s_i, s_j)}{\max(\sum_{j=1}^n c(s_k, s_j))}, \quad i \neq j. \quad (9)$$

3.3 Sentence Scoring and Selection

In order to select sentences to produce summaries, as various works did in the past, we interpret the Extractive Summarization problem as a binary classification one, where the model is trained to decide between exclusion and inclusion of individual sentences in the summary.

The procedure consists in minimizing the binary loss in a supervised learning approach where the sentences' feature vectors and the labels that classify whether they are present or not in their respective document summaries are used to train the model.

After training, we use the model's ability to distinguish between "important" and "unimportant" sentences to, given an input document, score and select appropriate sentences among the others. The scoring process consists in obtaining the probability of a sentence being part of the summary given its feature vector x , for each of the sentences in the document. Then, similarly to (Nallapati et al., 2017; Kedzie et al., 2018; Xiao and Carenini, 2019), the document summary is obtained by ranking and selecting top sentences using that probability, respecting the length compression limit.

4 EXPERIMENTS

4.1 Datasets

In this work, we compare EBM and GAMI-Net models with other approaches on two public text summarization datasets, namely CNN/Dailymail (Hermann et al., 2015; See et al., 2017) and Pubmed (Cohan et al., 2018). The CNN/Dailymail summarization dataset consists of pairs of news articles and their main highlights constituting 312K documents. This dataset have been widely adopted on recent ATS works, especially by Recurrent Neural Network and Transformer-based approaches. We use the non-anonymized version of this dataset (See et al., 2017).

The PubMed dataset is a collection of scientific papers totaling 133K documents in which the abstract section is used as the summary references. This dataset have been used for evaluating long document summarization approaches as both document and summaries are usually longer than in news datasets (Xiao and Carenini, 2019).

As mentioned in Section 3, our approach uses individual sentences as the input instances for training. Considering that initially both datasets only possess abstractive summaries, we needed extractive oracle labels to train our models. Recently, different authors obtained those labels using automatic heuristics while working with these datasets (Nallapati et al., 2017; Kedzie et al., 2018; Liu, 2019; Xiao and Carenini, 2019), which we adopt here. For CNN/Dailymail, we generated labels using the scripts provided by (Liu, 2019)¹, and, for Pubmed, we utilized labels extracted and made public by (Xiao and Carenini, 2019)². Moreover, we adopted random undersampling to handle label imbalance during the training step.

4.2 Model Comparison

We compare EBM and GAMI-Net with some recent baselines, most of which produced by deep neural architectures, in the sense of outlining the summarization ability in contrast to these models, despite the notable differences in terms of interpretability. Tables 1 and 2 present the results and reporting authors in CNN/Dailymail and Pubmed datasets, respectively.

Additionally, we compare EBM and GAMI-Net models with other supervised machine learning classifiers, i.e., Logistic Regression (LR), Random Forest (RF) and XGBoost, using the exact same fashion and features described in Section 3. Each technique was trained and tested ten times and the average scores are considered for comparison purposes.

Overall approaches are evaluated using the ROUGE score metric (Lin, 2004) considering its broad adoption for extractive summarization systems. Also, we evaluate the sentence selection ability of the supervised classifiers computing summary F1 scores based on the oracle labels. Moreover, Lead baseline correspond to selecting the first sentences present in the documents as the summaries (respecting each dataset summary-length limit) and Oracle denote the pre-obtained oracle labels' scores. Our ROUGE scores were obtained using pyrouge³, a python wrap-

¹<https://github.com/nlpyang/BertSum>

²https://github.com/Wendy-Xiao/Extsumm_local_global_context

³<https://pypi.org/project/pyrouge/>

Table 1: Results on CNN/Dailymail dataset.

Models	Type	Interpretability	ROUGE F (%)			F1 (%)
			R-1	R-2	R-L	
Lead	–	–	40.12	17.54	36.30	–
Oracle	–	–	56.09	33.67	52.21	–
P-Gen (See et al., 2017)	Abs.	Low	39.53	17.28	36.38	–
BART + RD (Wu et al., 2021)	Abs.	Low	44.51	21.58	41.24	–
SummaRuNNer (Ghodratnama et al., 2020)	Ext.	Low	39.90	16.30	35.10	–
MatchSum (Zhong et al., 2020)	Ext.	Low	44.41	20.86	40.55	–
ExDoS (Ghodratnama et al., 2020)	Ext.	High	42.1	18.9	37.7	–
LR	Ext.	High	38.51	16.66	34.74	31.96
RF	Ext.	Low	39.46	17.34	35.71	32.71
XGBoost	Ext.	Low	39.58	17.50	35.82	33.52
EBM	Ext.	High	39.48	17.40	35.74	33.40
GAMI-Net	Ext.	High	39.52	17.42	35.76	33.40

Table 2: Results on PubMed dataset.

Models	Type	Interpretability	ROUGE F (%)			F1 (%)
			R-1	R-2	R-L	
Lead	–	–	37.38	12.65	33.71	–
Oracle	–	–	55.37	26.31	49.07	–
Discourse-aware (Cohan et al., 2018)	Abs.	Low	38.93	15.37	35.21	–
SummaRuNNer (Xiao and Carenini, 2019)	Ext.	Low	43.89	18.78	30.36	–
ES-LG (Xiao and Carenini, 2020)	Ext.	Low	45.18	20.20	40.72	–
ES-LG+MMR-S+ (Xiao and Carenini, 2020)	Ext.	Low	45.39	20.37	40.99	–
LR	Ext.	High	38.07	12.70	32.94	27.61
RF	Ext.	Low	40.16	14.13	34.98	30.26
XGBoost	Ext.	Low	40.16	14.18	34.97	30.86
EBM	Ext.	High	39.86	13.96	34.65	30.70
GAMI-Net	Ext.	High	39.78	13.92	34.57	30.76

per of the original ROUGE-1.5.5 scripts and, concerning summary sizes, CNN/Dailymail summaries were limited to three sentences (Zhong et al., 2020) while Pubmed summaries were limited to 200 words (Xiao and Carenini, 2019).

4.3 Discussion

As denoted in Tables 1 and 2, EBM and GAMI-Net achieved similar results on both datasets. Considering ROUGE scores, GAMI-Net is ahead on CNN/Dailymail while EBM is superior on Pubmed, for less than 0.1 for each variant.

As shown in Table 1, comparing to other approaches, EBM and GAMI-Net models were able to compete with SummaRunner (Nallapati et al., 2017) and Pointer-Generator (See et al., 2017) networks – two of the earliest deep summarization architectures proposed in the past, surpassing the former concerning R-L and both of them considering R-2 on the CNN/Dailymail dataset. On the other hand, they

could not overcome BART+RD (Wu et al., 2021) and MatchSum (Zhong et al., 2020) (Transformer-based models) or ExDoS (Ghodratnama et al., 2020) in terms of scores. On the Pubmed dataset, as Table 2 shows, EBM and GAMI-Net achieved higher R-L scores than SummaRuNNer, but failed to compete with ExtSum-LG (Xiao and Carenini, 2019) and ExtSum-LG+MMR-S+ (Xiao and Carenini, 2020).

Although EBM and GAMI-Net fail to approximate the ROUGE scores of the most advanced summarization approaches, they are able to provide higher levels of transparency on how predictions are being built. For Extractive Summarization, this could be useful for clarifying what is being considered by the models while deciding the importance of the sentences. Figure 1 presents the plot of shape functions built upon the *Position* feature (Equation 5) given the respective model and dataset, where the horizontal axes represent feature values and the vertical axes denotes the corresponding shape function outputs. In practice, this kind of view allows further investigation

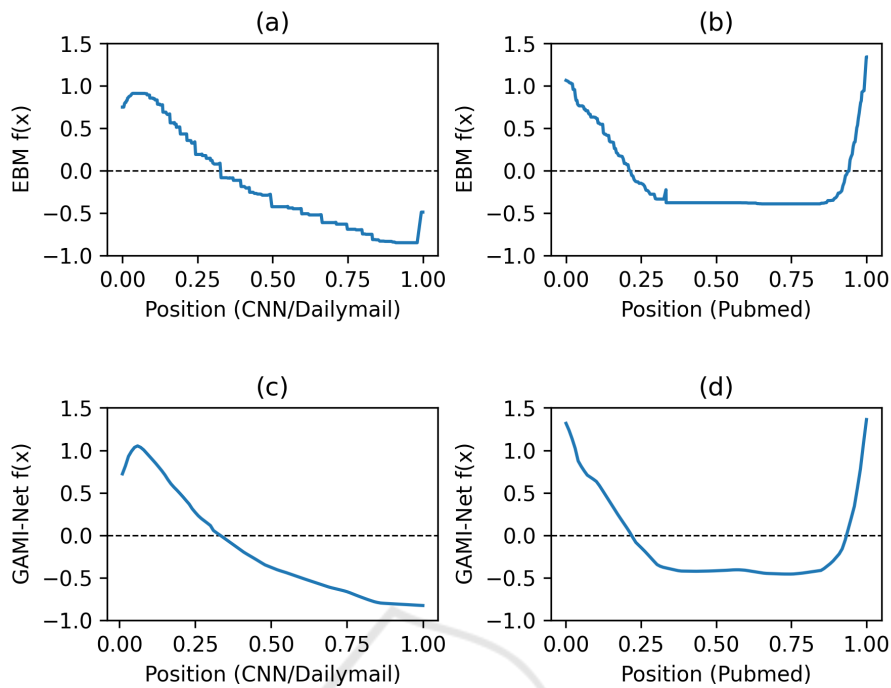


Figure 1: *Position* shape function.

of the summarization models, for example, inspecting how training models on different document types may be affecting the learning.

As Figure 1 shows, the shape functions fit on CNN/Dailymail (Figures 1a and 1c) tend to give higher outputs to sentences in the beginning of the documents, while on Pubmed (Figures 1b and 1d) sentences both at the beginning and the end of the documents are likely to be prioritized by the models. In this case, a possible assumption is that models trained on CNN/Dailymail (news type) are likely to avoid sentences at the end of the documents when looking for relevant sentences, which is not the case for all types of documents, as seems to happen for models trained on Pubmed (scientific article type/long document summarization). EBM and GAMI-Net ability to capture such properties intrinsically and grant the possibility of further exploration can be seen as a great benefit when comparing to full-complexity models.

Additionally, EBM and GAMI-Net feature outputs can be easily investigated for single predictions or a group of samples, helping better understanding of feature contributions while producing the summaries. Figures 2 and 3 show the most contributive feature effects quantified by their variation (Yang et al., 2021) on test set considering CNN/Dailymail and Pubmed datasets, respectively, where *Sentence-Sentence Similarity*, *TF-ISF* and *Position* shape functions are, in general, prevailing in importance over the others.

Comparing to other binary classifiers, both EBM and GAMI-Net placed in between XGBoost and Logistic Regression models, reinforcing their position as a balance between predictive power and transparency.

A key limitation of using GAMs with interactions for extractive summarization is the absence of a

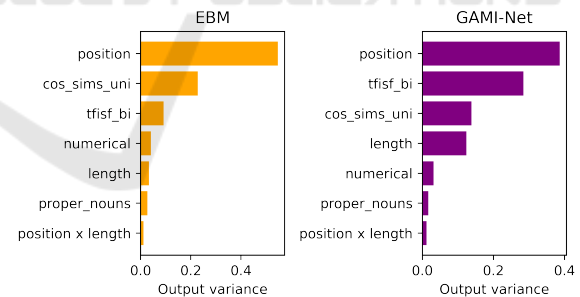


Figure 2: Top-7 importance ratios on CNN/Dailymail dataset.

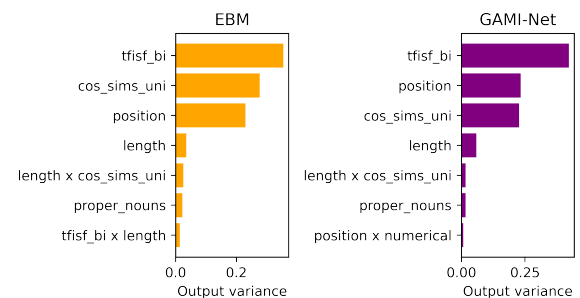


Figure 3: Top-7 importance ratios on Pubmed dataset.

mechanism to select good “sets” of sentences, rather than just individually relevant sentences. Furthermore, the semantic information gap, arising from the challenge of incorporating it through non-dense features, could be a means of obtaining even more powerful models in the future.

5 CONCLUSIONS

In this work, we present the application of EBM and GAMI-Net to interpretable extractive summarization, as a simple but attractive alternative to traditional classification algorithms. Our results show that, despite more restrictive than full-complexity models in terms of formulation, GAMs with interactions were able to achieve similar results to former black-box models.

Although the need for feature engineering can be seen as a disadvantage when comparing traditional approaches to neural models, with a concise set of features, both EBM and GAMI-Net models showed promising results for extractive summarization in textual datasets. The combination of intelligible features and the transparency of GAMs with interactions can be a tool to enlighten the view of the extractive summarization decisive process.

We present this paper as a preliminary effort concerning the topic of learning-based interpretable extractive summarization and believe that the perceptions presented into this work could help future research exploring the topic of intelligibility for ATS systems.

ACKNOWLEDGEMENTS

The authors are grateful to FAPESP grants #2013/07375-0, #2014/12236-1, #2019/07665-4, #2019/18287-0, and #2021/05516-1, and CNPq grant 308529/2021-9.

REFERENCES

- Afsharizadeh, M., Ebrahimpour-Komleh, H., and Bagheri, A. (2018). Query-oriented text summarization using sentence extraction technique. In *2018 4th international conference on web research (ICWR)*, pages 128–132. IEEE.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.
- Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O’Reilly Media, Inc.”.
- Cohan, A., Deroncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. (2018). A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621.
- Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., and Sen, P. (2020). A survey of the state of explainable ai for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459.
- Došilović, F. K., Brčić, M., and Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE.
- El-Kassas, W. S., Salama, C. R., Rafea, A. A., and Mohamed, H. K. (2020). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.
- Ferreira, R., de Souza Cabral, L., Lins, R. D., e Silva, G. P., Freitas, F., Cavalcanti, G. D., Lima, R., Simske, S. J., and Favaro, L. (2013). Assessing sentence scoring techniques for extractive text summarization. *Expert systems with applications*, 40(14):5755–5764.
- Ghodratnama, S., Beheshti, A., Zakershahra, M., and Sobhanmanesh, F. (2020). Extractive document summarization based on dynamic feature space mapping. *IEEE Access*, 8:139084–139095.
- Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386.
- Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.
- Kedzie, C., Mckeown, K., and Daumé III, H. (2018). Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828.
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Liu, Y. (2019). Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.
- Lou, Y., Caruana, R., and Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158.

- Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631.
- Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.
- Maybury, M. T. (1995). Generating summaries from event data. *Information Processing & Management*, 31(5):735–751.
- Montani, I., Honnibal, M., Honnibal, M., Landeghem, S. V., and Boyd, A. (2021). spaCy: industrial-strength natural language processing in Python.
- Moratanch, N. and Chitrakala, S. (2017). A survey on extractive text summarization. In *2017 international conference on computer, communication and signal processing (ICCCSP)*, pages 1–6. IEEE.
- Mutlu, B., Sezer, E. A., and Akcayol, M. A. (2019). Multi-document extractive text summarization: A comparative assessment on features. *Knowledge-Based Systems*, 183:104848.
- Nallapati, R., Zhai, F., and Zhou, B. (2017). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17*, page 3075–3081. AAAI Press.
- Nenkova, A. and McKeown, K. (2011). *Automatic summarization*. Now Publishers Inc.
- Nenkova, A. and McKeown, K. (2012). A survey of text summarization techniques. In *Mining text data*, pages 43–76. Springer.
- Nori, H., Jenkins, S., Koch, P., and Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.
- Oliveira, H., Ferreira, R., Lima, R., Lins, R. D., Freitas, F., Riss, M., and Simske, S. J. (2016). Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. *Expert Systems with Applications*, 65:68–86.
- Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K.-R. (2020). Toward interpretable machine learning: Transparent deep neural networks and beyond. *arXiv preprint arXiv:2003.07631*.
- Samek, W. and Müller, K.-R. (2019). Towards explainable artificial intelligence. In *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22. Springer.
- Sarkhel, R., Keymanesh, M., Nandi, A., and Parthasarathy, S. (2020). Interpretable multi-headed attention for abstractive summarization at controllable lengths. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6871–6882.
- See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.
- Wong, K.-F., Wu, M., and Li, W. (2008). Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd international conference on computational linguistics (Coling 2008)*, pages 985–992.
- Wu, L., Li, J., Wang, Y., Meng, Q., Qin, T., Chen, W., Zhang, M., Liu, T.-Y., et al. (2021). R-drop: regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34.
- Xiao, W. and Carenini, G. (2019). Extractive summarization of long documents by combining global and local context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3011–3021.
- Xiao, W. and Carenini, G. (2020). Systematically exploring redundancy reduction in summarizing long documents. *arXiv preprint arXiv:2012.00052*.
- Yang, Z., Zhang, A., and Sudjianto, A. (2021). Gami-net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognition*, 120:108192.
- Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., and Huang, X.-J. (2020). Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208.