Machine Learning Final Report

Faheem Shehzad Choudhry
26100036@lums.edu.pk
LUMS

Muhammad Abubakar Mughal
26100228@lums.edu.pk
LUMS

Muhammad Salman
26100177@lums.edu.pk
LUMS

M. Tayyab Haider
26100275@lums.edu.pk
LUMS

Ahmad Jawwad
25100237@lums.edu.pk
LUMS

ABSTRACT

This study addresses the challenges posed by the lack of computational resources for processing and categorizing Urdu text, a language significantly underrepresented in natural language processing research. To improve access to organized Urdu content we have developed a classification system which caters specifically to Urdu news articles. To create this model, we used a dataset consisting of 1000+ news articles from various news websites, which we then processed and cleaned to create the models. The articles were categorized into predefined classes, including entertainment, business, sports, science-technology, and international news. After studying evaluation metrics we found that the neural network achieved the highest accuracy, followed by the multinomial model and the logistic regression model. Our findings highlight the potential for improving accessibility to organized content in under-represented languages such as Urdu. Future work based on this research could expand upon the generalizability of the model.

KEYWORDS

Urdu, Classification, Multinomial, Neural Networks, Logistic Regression, Machine Learning, Web Scraping, Data Analysis, NLP, Transparency

ACM Reference Format:
Faheem Shehzad Choudhry, Muhammad Abubakar Mughal, Muhammad Salman, M. Tayyab Haider, and Ahmad Jawwad. 2024. Machine Learning Final Report. In *Proceedings of (*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Urdu has a limited presence in the field of NLP, which is particularly true within the scope of news articles categorization. In this regard, this project aims to help address the issue by creating a machine learning based classification model for Urdu news articles. Given that there was a lack of datasets, we collected more than 1000 news articles from local Urdu-speaking websites, including: Geo Urdu, Jang, and Express News. The said news pieces were then assigned to the following 'labels': Entertainment, Business, Sports, Science-Technology, and International.

The intended outcomes of the project also aimed at creating a good dataset, obtaining proper cleaning, and preprocessing the data and creating multiple machine learning models for accurate classification of the articles. Comparing and contrasting the effectiveness of 3 models: Logistic Regression, Neural Networks and Multinomial Naive Bayes. In doing this, it was hoped to try and determine which algorithms best suited the task. Each model was designed and implemented to the specific needs required in dealing with the problems of sparse text data, multi-lingual, and multidimensional.

We believe our work will benefit the evolution of tailored news systems for Urdu, which, in turn, should enhance the timely delivery of the right news personalized for the right user. This project has also shown how a combination of strong data preprocessing approaches and good modeling choices are critical for successfully addressing NLP problems in under-resourced languages.

2 METHODOLOGY

This section outlines the systematic approach we adopted to develop and evaluate models for classifying Urdu news articles. The methodology is structured into three phases, including data collection, preprocessing, and model implementation.

2.1 Data Collection

To develop a robust classification system for Urdu news articles, we collected a diverse dataset by scraping articles from multiple prominent Urdu news websites, including *Geo News*, *Express News*, and *Jang News*. These platforms were chosen for their popularity and coverage across a wide range of categories. By incorporating data from various sources, we aimed to ensure the dataset's generalizability and relevance across different topics and writing styles.

2.1.1 Rationale for Multi-Source Data Collection. Using multiple websites allowed us to capture linguistic and stylistic diversity inherent in Urdu journalism. Each platform targets slightly different audiences and editorial approaches, which helped create a dataset representative of the broader spectrum of Urdu news. This variety enhanced the robustness of our trained models, ensuring they perform well on unseen data and across varied contexts.

- **Targeted Category Scraping:** Specific categories were prioritized to align with the predefined labels for classification, more specifically we chose *Entertainment*, *Business*, and *Sports*. This ensured a balanced dataset across all categories.

978-1-4503-XXXX-X/18/06,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- **Data Structure:** For each article, we captured the following metadata:
  - Title
  - Content
  - Category
  - Source

2.2 Data Cleaning

The raw data we collected from various news websites contained significant noise and inconsistencies. This section outlines the steps undertaken to pre-process the dataset.

2.2.1 Handling Missing and Duplicate Values. Initial inspection revealed several instances of missing values, particularly in the *content* and *title* fields which occurred due to scraping errors. These records were removed to prevent biases during model training. In some instances we found duplicate articles, these were removed from the dataset as well.

2.2.2 Stopword Removal. To enhance the quality of the textual data, stopwords were removed from the *content* field. Since no existing Urdu stopword list was available in popular libraries such as NLTK, we created our own list of stopwords tailored to our Urdu language. This step helped our model focus on meaningful terms for classification.

2.2.3 Text Standardization. The textual data was standardized to retain only Urdu characters, ensuring consistency in formatting. Non-Urdu characters, hyperlinks, and excessive whitespace were removed to create a clean and uniform representation of the content.

2.3 Model 1: Logistic regression

The first model implemented was logistic regression, a baseline approach to classification tasks. The model was trained on the processed data set and evaluated using standard metrics such as accuracy, precision, and recall.

Motivation to use Logistic regression

Logistic regression is one of the very commonly applied models for classification problems. That is because it easily copes with text-based data in a very efficient way. It is an algorithm to estimate probabilities with the aid of a linear decision boundary, making it less difficult to implement and interpret in comparison to other complex models. In the model, the BOW approach(Bag of Words) was followed wherein articles were stacked vertically and arranged as high dimensional sparse word count-based matrix. Logistic regressions can deal with sparse matrices, as they use a weighted linear combination of features to compute decision boundaries, focusing on non-zero feature values when making predictions.

Furthermore, another critical factor, which motivates the choice of Logistic Regression is the assumption that each feature contributes independently to the target variable. This aligns with our underlying assumption of linear independence, where the occurrence of each word is only weakly related to the occurrence of the other.

Methodology

The first step was to transform the preprocessed text into its numerical representation using the BoW model. Using the CountVectoriser from the sklearn library, the model produced a sparse matrix where each row corresponded to an article and each column represented the frequency of unique words. This captured the textual features necessary for the classification.

A custom implementation of Logistic Regression was developed to classify news articles into their respective categories. The use of the One-vs-All (OvA), also sometimes referred to as One-vs-Rest(OvR), breaks down the multiclass classification problem into several binary classification problems, one for each class. For k binary classifiers we train k separate binary logistic regression models. Before each training, each class is separated from all the others by assigning the value of 1 for the target class, and the value of 0 for all the other classes. When making a prediction, the model will calculate the probability of a subset belonging to a respective class, using the sigmoid function. The class with the highest probability is assigned as the final prediction class.

The training process involved minimizing the cross-entropy loss using a gradient descent algorithm. In Gradient Descent, at each step of the iteration, gradients of the loss function are computed with respect to the model parameters, weight, and bias. The calculated gradients *dw & db* are used to update the weights and biases respectively. The learning rate is set as an important parameter, carefully tuned to balance between convergence and stability.

In the context of text-classification techniques, the matrices often contain a significant number of zero values. This occurs because most words do not appear in every document. This results in slower computation times due to the density of the high-dimensional matrix. To counter this, the use of *csr_matrix* is a crucial step in converting data to sparse matrices. Not only does this approach save memory, but it also accelerates computations by ensuring that operations are performed only on the non-zero values. This makes the model scalable, allowing it to handle larger datasets with limited computational resources.

Evaluation and Performance Analysis

The classification results demonstrate strong model performance on both the training and test datasets, with an overall train accuracy of 98% and a test accuracy of 95%. Precision, Recall and F1-Scores remain high across all categories, indicating that the model is quite effective at separating different classes. However some performance variations are observed between the training and test datasets. An example can be where the business category performs well with the training set, showing an F1 score of 0.98 on the training set, but that drops to 0.93 on the test set, suggesting overfiting. Table 1 shows the results, demonstrating the Precision, Recall, F1-Score and Support for each class. In comparison, the "science-technology" category has 99% recall, which points to a high ability of the model to recognize instances from this class correctly.

The consfusion matrix in Table 2 demonstrates the number of correct classifications The confusion matrix of the training set shows very few errors, with only some scattered ones in different categories, like four misclassified "science-technology" and two in "world."The confusion matrix for the test set shows a very specific pattern of missclassifying categories. For instance, three "business" instances were classified as "science-technology," and one "science-technology" was misclassified as "world." However, the model performs very well in both "sports" and "entertainment," with virtually perfect precision and recall.

For Further improvement in the accuracy of the Logistic Regression model on test data, regularisation techniques, such as Lasso(L1

Regularisation) and the Ridge(L2 Regularisation), can be used to reduce overfitting. These techniques help in reducing overfitting and improving generalization to unseen test data. Ridge adds a penalty proportional to the square of the coefficients, which helps in shrinking coefficients of less useful features. However, Ridge does not reduce the coefficients to zero, thus does not perform feature selection (Murel, 2023). While, Lasso, applies a penalty proportional to the absolute value of the coefficients; which actually leads to feature selection and simplification of the model. Repeating the classification with Ridge and Lasso, we obtained an accuracy of 95% and 94% on training data respectively, indicating little to no change. The failure to show any improvement with Lasso or Ridge regularisation indicates that the model and dataset are well-suited to the model, without any additional constriants.

Training Classification Report				
Category	**Precision**	**Recall**	**F1-score**	**Support**
Business	1.00	0.97	0.98	183
Entertainment	0.99	0.99	0.99	183
Science-Technology	0.96	0.99	0.98	182
Sports	0.99	0.99	0.99	184
World	0.97	0.98	0.98	183
Accuracy	0.98 (915)			
Macro avg	0.98	0.98	0.98	915
Weighted avg	0.98	0.98	0.98	915
Test Classification Report				
Category	**Precision**	**Recall**	**F1-score**	**Support**
Business	0.97	0.89	0.93	44
Entertainment	0.98	0.96	0.97	46
Science-Technology	0.90	0.96	0.93	47
Sports	1.00	0.98	0.99	45
World	0.92	0.98	0.95	47
Accuracy	0.95 (229)			
Macro avg	0.95	0.95	0.95	229
Weighted avg	0.95	0.95	0.95	229

Table 1: Training and Test Classification Report

True \Predicted	**B**	**E**	**ST**	**S**	**W**
B	39	0	3	0	2
E	0	44	1	0	1
ST	1	0	45	0	1
S	0	1	0	44	0
W	0	0	1	0	46

Table 2: Confusion Matrix for Test Set

Category Descriptions:

- **B**: Business
- **E**: Entertainment
- **ST**: Science Technology
- **S**: Sports
- **W**: World

True \Predicted	**B**	**E**	**ST**	**S**	**W**
B	177	0	4	1	1
E	0	181	0	0	2
ST	0	0	180	0	2
S	0	1	0	183	0
W	0	0	3	0	180

Table 3: Confusion Matrix for Training Set

2.4 Model 2: Neural Network

The second model implemented for this project was a neural network, which leverages multiple layers and activation functions to capture non-linear patterns in the dataset. This model demonstrated exceptional performance and was the highest-performing model in our study.

2.4.1 Motivation for Using a Neural Network. Neural networks were chosen for this project due to their ability to capture complex, non-linear relationships in data, making them particularly suited for tasks like text classification where features are highly interdependent. Traditional methods like Logistic Regression and Naive Bayes rely on linear relationships and independent feature assumptions, which may limit their effectiveness when dealing with rich, high-dimensional text data. Neural networks, on the other hand, can model intricate patterns and relationships between words, allowing for better generalization and improved performance, especially in more complex and diverse datasets like the Urdu news articles used in this project. The flexibility of neural networks also allows for better adaptation to varied linguistic features, which is crucial for handling the nuanced nature of text in different languages, including Urdu.

2.4.2 Model Architecture. The neural network consists of a three-layer fully connected architecture designed for multi-class classification of Urdu articles. The structure of the model is as follows:

- **Input Layer:** A fully connected layer with 128 units.
- **Hidden Layer:** A fully connected layer with 32 units, utilizing the ReLU activation function.
- **Output Layer:** A fully connected layer with a number of units equal to the number of categories, followed by softmax activation to compute probabilities for each label (class).

This architecture processes feature vectors extracted from text data, enabling effective classification into multiple categories.

2.4.3 Performance and Strengths. The neural network achieved the highest accuracy among all tested models, with a peak test set accuracy of 97.4%. Key performance highlights include:

- **High Accuracy:** Training accuracy reached over 99%, while validation accuracy consistently hovered around 96–97%, indicating the absence of overfitting.
- **Trend Capture:** The model's architecture enabled it to quickly learn trends in the dataset. Each neuron identified specific features and nuances, which contributed to its strong performance.
- **Efficient Training:** The validation peak accuracy was achieved within the first 5 epochs out of the total 20 epochs, demonstrating training efficiency.

2.4.4 Limitations.

- **Overfitting Risks:** Despite its strong performance, the neural network was highly susceptible to overfitting when the architecture was made more complex, such as by increasing the number of layers or neurons per layer.
- **Sensitivity to Parameters:** The balance between training efficiency and validation accuracy was delicate. Minor changes in model parameters led to significant degradation in generalization performance.
- **Computational Complexity:** The model, while efficient for this task, requires more computational resources compared to simpler models like Multinomial Naïve Bayes.

2.4.5 Conclusion. Overall, the neural network proved to be a highly effective model for the classification of Urdu articles, achieving state-of-the-art performance on this dataset. While its susceptibility to overfitting and computational demands present challenges, the carefully designed architecture demonstrated that neural networks can achieve remarkable accuracy and generalization in text classification tasks.

2.5 Model 3: Multinomial Naïve Bayes (MNB)

The Multinomial Naïve Bayes (MNB) model proved to be a strong candidate for the classification of Urdu news articles, balancing simplicity, interpretability, and high performance. Since it is designed specifically for text-classification tasks, the model was perfectly suited for the task at hand.

Some specific reasons we found the model to be particularly effective was that:

- **Category-Specific Characteristics:** The categories in our dataset (e.g., *Entertainment*, *Sports*, *Business*) often have distinct vocabularies. For instance, the *Entertainment* category frequently mentions celebrity names and film titles, while *Sports* articles contain words like "match" or "team." The model was able to capitalize on these frequency differences to make accurate predictions.
- **Simplicity and Efficiency:** The model is computationally efficient unlike other models where we would have to repeatedly adjust hyperparamters over multiple iterations. This made it ideal for tasks where performance is an important factor.
- **High Performance:** Despite its simplicity, the MNB achieved a remarkably high accuracy of 96.4%, comparable to more complex models like neural networks, which demonstrated its suitability for this task.

2.5.1 Performance and Strengths. The model performed well in several different ways:

- **High Accuracy:** The overall accuracy of the model was 96.4%, which outperformed our expectations as it came within 0.5% of the neural network's accuracy, making it one of the best-performing models in this study.
- **Transparency and Interpretability:** MNB's probabilistic foundation allowed for great simulatability as its simple underlying mechanism allowed for easy understanding
- **Category-Level Performance:** The model demonstrated high precision and recall across all categories, as reflected in its F1-scores:
  - **Business:** F1-score of 0.99.
  - **Entertainment:** F1-score of 0.97.
  - **Science and Technology:** F1-score of 0.96.
  - **Sports:** F1-score of 0.99.
  - **World:** F1-score of 0.91.

2.5.2 Limitations.

- **Independence Assumption:** The assumption that words are independent given the class, while useful, may not always hold true. For example, certain word combinations (e.g., "World Cup" in *Sports*) may carry more contextual meaning than individual words.
- **Handling of Rare Words:** Rare words that appear infrequently in the dataset may not be as effectively utilized, as MNB relies heavily on word frequency distributions.
- **Model Scalability:** While MNB works well for a moderate-sized dataset like ours, its simplicity may limit its performance on larger, more complex datasets with richer linguistic features.

2.5.3 Conclusion. Overall, we found that the Multinomial Naïve Bayes model emerged as the best choice for this text classification task. Its combination of simplicity, efficiency, and transparency made it a compelling option. Given its high accuracy and interpretability, MNB served as a valuable benchmark and demonstrated that simple systems can yield competitive results for classifying Urdu news articles.

3 FINDINGS

This project explored three models for classifying Urdu news articles into multiple categories, leveraging a diverse dataset collected from prominent news websites such as Geo News, Express News, and Jang News. By integrating data from these sources, the dataset captured a wide range of linguistic and stylistic variations, enhancing the generalizability and robustness of the models.

3.1 Logistic Regression

- **Performance:** Achieved a training accuracy of 98% and a test accuracy of 95%, with high precision, recall, and F1-scores across all categories.
- **Strengths:**
  - Efficient for handling high-dimensional sparse data using the Bag-of-Words (BoW) approach.
  - Simple and interpretable, leveraging the One-vs-All (OvA) strategy for multiclass classification.
  - Sparse matrix optimization improved computation speed and memory usage, making it scalable for larger datasets.
- **Limitations:**
  - Slight overfitting observed in some categories, such as *Business*, where the F1-score dropped from 0.98 (train) to 0.93 (test).
  - Did not benefit significantly from additional regularization techniques (L1/L2), indicating limited scope for further optimization with the current dataset.

- Limited ability to capture non-linear patterns compared to more advanced models like neural networks.

3.2 Neural Network

- **Performance:** Achieved the highest overall performance, with a peak test set accuracy of 97.4%.
- **Strengths:**
  - Captured non-linear patterns effectively through its three-layer architecture.
  - Demonstrated strong generalization with minimal overfitting.
  - Quickly achieved peak performance within 5 epochs.
- **Limitations:**
  - Susceptible to overfitting when the architecture was made more complex.
  - More computationally intensive compared to simpler models like Naïve Bayes.

3.3 Multinomial Naïve Bayes (MNB)

- **Performance:** Delivered remarkable accuracy of 96.4%, comparable to the neural network.
- **Strengths:**
  - Capitalized on category-specific vocabularies to make accurate predictions.
  - Simple, efficient, and highly interpretable.
  - Achieved high precision, recall, and F1-scores across all categories.
- **Limitations:**
  - Assumes word independence, which may not always hold true for contextual word combinations.
  - Struggles with rare words due to reliance on word frequency.
  - Simplicity may limit scalability to larger, more complex datasets.

3.4 Impact of Data Selection on Model Performance

Ensuring that the data for these models was selected from a wide range of sources and included diverse topics was key to generalize the performance obtained from our models. Furthermore, the process of cleaning and preparing the data also impacted results.

- **Diverse Sources:** Collecting news articles from multiple prominent platforms such as Geo News, Express News, and Jang News ensured linguistic and stylistic diversity. This variety allowed the models to generalize better to unseen data and different contexts.
- **Category Balance:** Ensuring that all labels (e.g., *Entertainment*, *Business*, *Sports*) received a fair share of representation in the dataset ensured a balanced distribution and prevented any biases from occurring in the data.
- **Data Cleaning:** Comprehensive cleaning techniques, including duplicate articles removal, replacement of missing values, standardization of text and stop-word removal ensured high-quality input for the models.

3.5 Key Takeaways

- The **Logistic Regression** model offered a solid baseline for performance evaluation.
- The **neural network** proved to be the most effective model, offering the best performance but at the cost of higher computational overheads as well as over-fitting risks.
- The **Multinomial Naïve Bayes** model, while simpler, offered competitive accuracy and efficiency. It was within 1% of the neural network making it a strong alternative for text classification tasks.
- The project demonstrated that combining robust data preprocessing with appropriate modeling choices can yield high-performance systems for Urdu text classification. Furthermore, the difference between complex computationally intensive models and more understandable linear models is surprisingly small.

4 LIMITATIONS

Despite achieving promising results, the project faced several limitations related to the dataset, model selection, and the inherent challenges of text classification using machine learning.

4.1 Data-Related Limitations

Combining Urdu text with the existing datasets resulted in a wide range of difficulties, including the scripting logistics as well as the use of context-sensitive words that are often ignored in normal text processing methods. The partitioning semantic's richness which is a Property of Urdu language is particularly challenging here because this leads to an increase in the complexity of the preprocessing phase. Also, even though category balancing was performed, some categories had slight overlaps which made things a bit more complicated for the model. This was particularly obvious in the "Business" category in which only slight drops in performance for the Logistic Regression model were noted. Also, the dataset was global but was limited in terms of the already available labelled Urdu news articles which reduced their potential in being generalized to new unpublished data. The custom Urdu stopword list while addressing the removing of irrelevant words likely also removed the features that revolve around the words in the text and thus losing the models' capturing of what is essential in the text.

4.2 Model-Specific Limitations

Each and every model used in this project had its own limitation. For instance, Logistic Regression assumes that relationships between features are linear which is very restrictive when it comes to the multivariate and non-linear interactions that are typical in text data systems. Especially for tasks requiring recognition of more complex patterns, this model is poor in performance due to the fact that the model cannot represent higher order word and phrase interactions. The Multinomial Naïve Bayes model on the contrary assumes independence of words thus oversimplifying the relationships as it hinders the model when it is faced with dependency and heading a contextual or phrasal information. It is to be noted that the result from the neural network was very good, however it consumed too much of computational power and overfitted the data, especially with a small size of the data set. As well as, the L1 and

L2 regularization methods didn't greatly improve the results across the Logistic Regression and the Neural Networks which indicates that the problem may be the variety or capacitance of the models feature.

4.3 Generalizability

The dataset used in this study was sourced exclusively from news websites, which limits the models' ability to generalize to other domains, such as social media or academic texts. This domain-specific bias restricts the applicability of the models outside of the news industry, and the models may struggle to perform as effectively on content with different linguistic structures or informal language. Additionally, the linguistic style and topical focus of formal news language evolves over time. This necessitates periodic retraining of the models to maintain their relevance and accuracy, particularly as new trends, topics, and vocabulary emerge in the media landscape.

5 FUTURE CONSIDERATIONS

After concluding this project, we have identified several key considerations and improvements that should be taken into account, whenever a similar project is set-up for text classification purposes in the future, to improve the overall results. Firstly, expanding the dataset size and making it diverse is essential for enhanced generalizability and robustness. Incorporating content from a broader range of sources, such as blogs, social media, and academic materials will ensure that it captures more varied linguistic styles and domain-specific content. Additionally, developing the ability to detect multi-class or overlapping content would address scenarios where articles span multiple categories which is commonplace when considering real-life applications, thereby improving results.

From a model-specific perspective, minor improvements to the existing implementation can significantly enhance performance. For logistic regression, more advanced regularization techniques like elastic net could be explored to better manage high-dimensional sparse data while mitigating over-fitting. Neural network architectures could benefit from additional layers, provided that the dataset has grown significantly, to better capture textual-context and semantics. For the most part, the main bottleneck for our Neural Network implementation was the limited dataset which greatly restricted our architecture complexity since it was super prone to over-fitting. This means that even a slight increase in the model complexity renders neural networks' overkill for tasks reliant on limited datasets. Multinomial Naïve Bayes, on the other hand, could be refined to handle word dependencies more effectively, such as by integrating n-gram approaches to account for contextual word combinations. These model-specific refinements, coupled with broader and more domain-rich data would pave the way for a more versatile and scalable text classification system.

6 CONCLUSION

This study has conclusively verified the effectiveness of machine learning approach in solving the problem of classifying Urdu language news articles. Out of the three models tested, the neural network had the best performance, demonstrating its ability to learn complicated structures of text data. However, the Multinomial Naive Bayes model turned out to be a good alternative due to its relatively good performance at a lower cost in terms of scope and complexity in implementation. Logistic Regression was also included in the experiments and provided a good benchmark as it performed well and remained moderate in terms of explanation and prediction.

At the heart of the project was the design of a good dataset because of the effort put into data collection and cleaning. Other methods such as text normalization, removing stopwords and scraping of specific categories were put in place to ensure the models would be able to transfer across multiple topics. Nonetheless, the effort was worthwhile since a number of limitations with this approach were evident; for example, risks of over-fitting in complicated models or relying on too strong independence assumptions in simple ones.

This work can be extended by bettering what has been done through studying more complex NLP methods like transformer-based models or expanding the dataset to include more categories. This paper optimistically presents promising possibilities for machine learning in organizing content for users speaking less available languages such as Urdu, which constitutes a first step towards other advances for content distribution systems.

ACKNOWLEDGMENTS

Special thanks to our assigned TA Danyal who verified all our models were working as intended and Sir Amin for answering our queries on slack.

REFERENCES

- Murel, J. (2023). What is ridge regression? | IBM. [online] Available at: https://www.ibm.com/topics/ridge-regression.
- Nam, J., Kim, J., Loza Mencía, E., Gurevych, I., and Fürnkranz, J. "Large-Scale Multi-Label Text Classification — Revisiting Neural Networks." *Machine Learning and Knowledge Discovery in Databases*, edited by T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, vol. 8725, Springer, 2014, pp. 437–452. *Lecture Notes in Computer Science*, https://doi.org/10.1007/978-3-662-44851-9_28.

978-1-4503-XXXX-X/18/06,