Literary Voice: *A Peer Reviewed Journal of English Studies (ISSN 2277-4521)* Number 25, Volume 1, September 2025, https://literaryvoice.in
Indexed in the Web of Science Core Collection ESCI, Cosmos, ESJI, I20R, CiteFactor, InfoBase

Predictive Analysis of Grimm's Fairy Tales: Leveraging Logistic Regression for Classification and Evaluation ♣

[1]**Dr. Aditi Goyal,** Assistant Professor, Dept. of English and Liberal Arts, Chandigarh University, Gharuan, Mohali, S.A.S. Nagar (Punjab), India. Email: draditigoyal.11@gmail.com
[2]**Dr. Vineet Mehan**, Professor, NIMS University, Jaipur (Rajasthan), India. mehanvineet@gmail.com
DOI: https://doi.org/10.59136/lv.2025.25.1.43a

Abstract

The research paper aims to discover automatic classification of Grimm's Fairy Tales by exploring the Machine Learning insights. A corpus of 44 tales is taken for the experimentation purpose. A set of 5 predictor variables are Title, Abstract, Content, Aarne-Thompson-Uther (ATU) Numerical and ATU Type. ATU topic is taken to be the response variable. Preprocessing techniques applied include Transformation, Tokenization and Filtering. Bag of Words model is applied to generate the numerical representation of contents in the form of a vector. Word cloud represents the existence of words on the basis of frequency of occurrence. Various regression algorithms are applied for predicting the Grimm's classification. Nomogram represents the relationship among top 4 variables identified for classification. Parameters for evaluation include Area under Curve (AUC), Classification Accuracy (CA), F1 Score, Precision, Recall and Matthews Correlation Coefficient (MCC). Effectiveness of machine learning model is achieved for automatic classification by evaluating the parameters and achieving correlations. The research. work displays the effectiveness of logistic regression in analysing and classifying Grimm's Fairy Tales, giving important insights into the predictive modelling of literary texts.
Keywords: Predictive Analysis; Logistic Regression; Grimm's Fairy Tales; Classification; Machine Learning

Introduction

In recent years, the integration of machine learning into literary analysis has gained momentum, offering new ways to explore, interpret, and classify texts. Among literary genres, fairy tales hold a unique place due to their cultural longevity, narrative structure, and symbolic richness. The Grimm Brothers' fairy tales, first published in the early 19th century, have been studied extensively from folkloristic, linguistic, and cultural perspectives. Early applications of machine learning to folklore often focused on identifying linguistic similarities across regional variations (Davis et al., 459).

These tales—often centred around moral lessons, archetypal characters, and fantastical elements—have inspired generations of readers, scholars, and artists. Yet, from a computational standpoint, they present an intriguing challenge: can a machine automatically classify them based on narrative or linguistic features?

Fairy tales share certain universal traits—such as recurring motifs, moral dichotomies, and formulaic openings or endings—that suggest an underlying "quintessence" to the genre. However, individual stories often diverge in tone, complexity, and narrative structure, making computational classification a non-trivial task. While natural language processing (NLP) has been widely applied to news articles, scientific texts, and contemporary media, its application

♣**Article History**: Full Article Received on 2nd June 2025. Peer Review completed on 30th July 2025, Article Accepted on 15th Aug 2025. First published: September 2025.

to fictional narratives—especially folk literature—remains comparatively underexplored. Previous research on computational approaches to storytelling has focused more on plot structure detection, sentiment analysis, or author style recognition, with limited work specifically targeting Grimm's fairy tales (Foster 45). This research seeks to address that gap by applying machine learning techniques—specifically logistic regression—to classify a curated corpus of forty-four Grimm's fairy tales. By examining linguistic, structural, and thematic features, the study aims to determine whether automated classification can effectively distinguish between stories within this corpus. The significance of this research lies not only in expanding the scope of computational literary analysis but also in offering insights into the fundamental nature of fairy tales. If successful, such an approach could contribute to broader applications in digital humanities, automated story categorization, and the preservation and analysis of folklore traditions.

Literature Survey

Grimm's fairy tales, originally compiled by Jacob and Wilhelm Grimm, have been central to folklore studies and literary scholarship for centuries. Scholars have analysed their moral, psychological, and cultural dimensions. In *The Hard Facts of the Grimms' Fairy Tales*, Maria Tatar explores the sociopolitical undertones and psychological depth of these stories, arguing that they reflect anxieties of the societies from which they emerged (Tatar). Jack Zipes, in *The Brothers Grimm: From Enchanted Forests to the Modern World*, investigates how the tales evolved across editions, revealing shifts in gender roles, violence, and nationalism, indicating the editorial choices made by the Grimms to align with contemporary values (Zipes). These perspectives lay the foundation for understanding the fairy tales not just as stories, but as texts rich in symbolic and thematic content that can be categorized and analysed.

Logistic regression is a foundational machine learning technique used for binary and multiclass classification problems. In text analysis, it is particularly effective when combined with feature extraction methods like TF-IDF (Term Frequency–Inverse Document Frequency). Jurafsky and Martin, in *Speech and Language Processing*, highlight the effectiveness of logistic regression for text classification, noting its interpretability and relatively low computational cost (Jurafsky and Martin). Similarly, Sebastiani's review on machine learning in automated text categorization underscores the relevance of logistic regression for classifying documents in various domains including sentiment analysis and genre detection (Sebastiani).

Recent works also explore the combination of literary analysis and computational methods. Jockers' *Macroanalysis* presents methods for analysing large literary corpora using statistical models, offering a precedent for using machine learning in literary studies (Jockers). This approach suggests the potential of logistic regression to classify fairy tales by features such as moral tone, protagonist type, or thematic structure. By Integrating literary theory with computational techniques, this research aims to identify and evaluate patterns across Grimm's Fairy tales using logistic regression, providing insights that traditional close reading might overlook.

Methodology

Various machine learning techniques for classification algorithms include Logistic Regression, k Nearest Neighbour (kNN), Decision Tree, Random Forest, Support Vector Machine (SVM) and Naive Bayes. By simulating the correlation between characteristics and the likelihood of falling into a specific class, the classification process known as logistic regression forecasts categorical results. It maps predictions to probabilities ranging from 0 to 1 using a logistic function, which makes it useful for linearly separable data and flexible for multi-class situations. In addition to being computationally economical, it may contain regularization to avoid overfitting and performs well with high-dimensional datasets. Common hyperparameters

include penalty, inverse of regularization strength, solver, maximum optimization steps, and class weight.

In logistic regression an S shaped curve called sigmoid function is used. Threshold value is for the classification purpose. If the outcome is above threshold, then one category is selected. If the outcome is below threshold, then another category is chosen. The algorithm can be modified from binary classification to multi-class classification.

A non-parametric technique called the k-Nearest Neighbors (kNN) algorithm uses the majority class of a data point's k nearest neighbors in the feature space to classify it. It doesn't assume anything about the distribution of the underlying data; instead, it uses a distance measure to assess similarity. Although kNN is easy to use and performs well on smaller datasets, the curse of dimensionality may cause it to lose some of its effectiveness when dealing with high-dimensional data. The number of neighbors to take into account, weights, the Euclidean metric, the power parameter for calculating distance, and the method are examples of common hyperparameters. Another algorithm for predicting the classification is kNN. The most common metric used for prediction in kNN is Euclidean Distance. Using this distance, we can find the class to which the new point will belong to. The point with the smallest distance is chosen as the reference class. k is the hyperparameter that determines the learning rate. k here depicts the number of neighbours. A decision tree is a supervised learning system that uses a hierarchical, tree-like structure to express decisions. Each leaf node represents a predicted class, and each inside node divides the data according to a certain feature value. It is very interpretable and efficiently manages both numerical and categorical data, but if regularization or pruning are not used, it may overfit. The criteria gini index, the maximum tree depth, the minimum number of samples needed to split a node, the minimum number of samples at a leaf node, and the number of features to take into account at each split are examples of common hyperparameters. In decision tree, a tree is made. In this tree, at each level decision is taken about the classification. Leaf nodes tell the final decision in this algorithm.

Alternatively, when a number of decision trees are taken then it becomes a random forest. In this tree observation is there for each decision tree. Majority voting is done to find out the final classification problem. In order to increase accuracy and decrease overfitting, Random Forest, an ensemble learning approach, constructs many decision trees and combines their predictions. To ensure model variety, it uses bootstrap aggregating to train each tree on a random sample of features and data. For classification problems, majority voting is used to make predictions, which makes the approach noise-resistant and able to handle big datasets with plenty of characteristics. The number of trees, gini index, maximum tree depth, minimum number of samples needed to divide a node, and maximum features are hyperparameters.

Another algorithm for classification is Naive Bayes. It uses Bayesian algorithm to calculate the probability. It is on the basis of this probability that data points are chosen belonging to a particular class and thus the classification.

Based on Bayes' theorem, the Naive Bayes family of probabilistic classifiers assumes that characteristics are conditionally independent given the class label. It is very useful for text classification jobs, is very efficient, and performs well with high-dimensional data. There are other variations that are specific to various kinds of data, such as Gaussian, Multinomial, and Bernoulli Naive Bayes. The alpha smoothing value for MultinomialNB and BernoulliNB to handle zero probabilities, whether to learn class prior probabilities, and manually setting prior probabilities are examples of common hyperparameters. GaussianNB is used to control numerical stability.

Experimental Results

For the experimental purpose, a corpus of 44 tales is taken as given in Table I. Each document within the corpus contains a target variable. Aarne-Thompson-Uther (ATU) Topic is the target variable in the document. There are 5 meta-attributes that include: Title, Abstract, Content,

ATU Numerical and ATU Type. The model used for experimentation purpose is demonstrated in Fig. I.

Table I: Corpus of Documents

Title of Documents					
A Tale About the Boy Who Went Forth to Learn What Fear Was	Mother Holle	The Crumbs on the Table	The Crumbs on the Table	The Queen Bee	The Willow-Wren and the Bear
Brier Rose	Old Sultan	The Dog and the Sparrow	The Dog and the Sparrow	The Raven	The Wolf and the Fox
Cat and Mouse in Partnership	Pack of Scoundrels	The Elves and the Shoemaker	The Elves and the Shoemaker	The Seven Ravens	The Wolf and the Man
Cinderella	Rapunzel	The Fisherman and His Wife	The Fisherman and His Wife	The Straw, the Coal, and the Bean	The Wolf and the Seven Young Kids
Hansel and Gretel	Rumpelstiltskin	The Fox and the Cat	The Fox and the Cat	The Three Languages	
Herr Korbes	Snow White	The Fox and the Geese	The Fox and the Geese	The Water of Life	
Jorinda and Jorindel	The Blue Light	The Fox and the Horse	The Fox and the Horse	The Wedding of Mr. Fox	
Little Red Riding Hood	The Bremen Town Musicians	The Frog Prince	The Frog Prince	The White Snake	

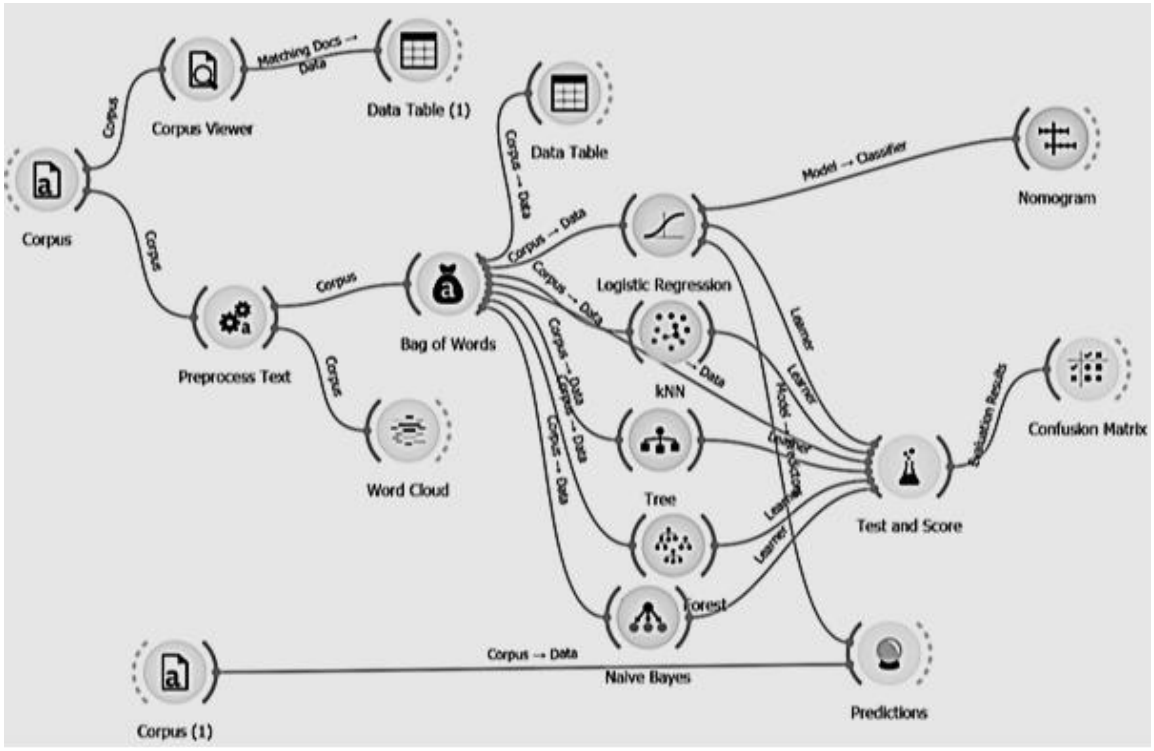

Fig. I. Experimentation Model

Preprocessing of the corpus is done in the first step. Three preprocessing techniques are applied in this process. These techniques are Transformation, Tokenization and Filtering. In Transformation all the words are reduced to lowercase. In Tokenization a regular expression "\w+" is used to match characters, numbers and underscore character. In filtering a few stop words are removed which include "could; would; said; should; came".
The preprocessing output is represented in the form of word cloud as given in Fig. II. The Word cloud is made on the basis of words and their weights in the corpus. The size of words gives information about the frequency of occurrence. The bigger the size, more is the frequency of words in corpus.

Fig. II. Word Cloud

Bag of words is a simple and computational efficient parameter which is used to convert words into vector. The vector is a numerical representation of words in the form of matrix. The matrix represents the frequency in terms of count of the words in a vocabulary. 3715 features are identified from all the 44 documents in corpus with a sparse density of 100%. Output of the parameter is further used for analysis with Machine Learning Algorithms.
In order to predict the Grimm's classification, various classification models are evaluated. Among all the classification model's Logistic regression outperformed all the models. The model is trained for the given corpus of 44 documents. A Nomogram of the top four variables identified for the classification is shown in Fig. III. Relationship between the variables and the predicted outcome can be visualized using the Nomogram.

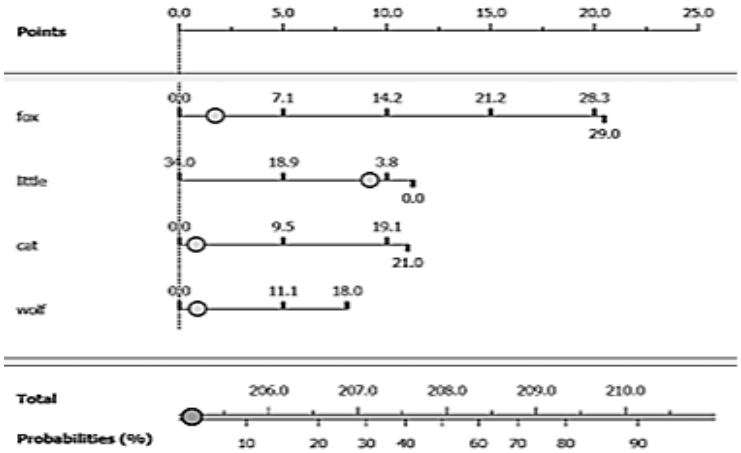

Fig. III. Nomogram

Figure III's nomogram offers a visual way to estimate probabilities based on a variety of category inputs, namely "fox," "little," "cat," and "wolf." Scales show the related point values for each variable, and blue circles indicate the chosen values for each category. For instance, "little" is equivalent to around 3.8 points, "cat" to 9.5 points, "wolf" to 11.1 points, and "fox" to roughly 7.1 points. The "Total Points" axis is used to transfer the total point score—which is obtained by adding these values—to a probability percentage at the nomogram's bottom. In this case, the total points add up to around 206, which corresponds to a probability that is just under 10%.

In order to get the results by testing and scoring the logistic regression, a stratified cross-validation approach is used with a fold of 5. Random sampling is used with a repeat train\test size of 10. The training set size of 66% is taken per iteration. Evaluation results for the target variable are given in Table II below.

Table II: Evaluation Results

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.945	0.88	0.881	0.89	0.88	0.764
kNN	0.851	0.667	0.652	0.818	0.667	0.492
Tree	0.771	0.76	0.762	0.766	0.76	0.51
Random Forest	0.923	0.787	0.788	0.824	0.787	0.607
Naive Bayes	0.941	0.487	0.395	0.775	0.487	0.252

Among all the classification algorithms an Area Under the Curve (AUC) value of 0.945 for Logistic Regression indicates very good performance for a classification model. It correctly classifies positive cases while maintaining a low negative cases. A Classification Accuracy (CA) of .880 depicts that the model is able to classify 88% of the data points. For identifying the overall percentage of correctly classified data points CA parameter is very useful. An F1 score of 0.881 specifies upright performance of the model's capability to categorise among positive and negative groups.

Precision of 0.89 tells that the proposed technique has a moderately low rate of false positives. Recall of 0.88, identifies that the proposed algorithm captured 88% of the actual positive cases. Matthews Correlation Coefficient (MCC) of 0.764 indicates a strong correlation between the true labels and the model's predictions.

The Confusion matrix of the predicted output to Actual values is depicted in Fig. IV. below. 132 entries are correctly classified out of a total value of 150. Only 18 values are misclassified out of 150.

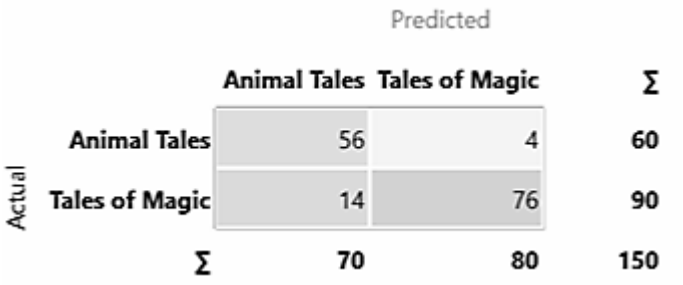

Fig. IV. Confusion Matrix

In order to identify the model's performance on unseen data, an Anderson corpus of 3 tales is taken. The trained logistic regression model is applied to this new data. For the title "The Little Match-Seller" and "The Philosophers' stone" the category identified came out to be "Tales of

Magic". While for the Title "The Ugly Duckling" the category identified was "Animal Tales". Similarly, a number of other literary texts can be classified using the proposed approach.
We made a feature selection modification to prevent bias and data leakage that might result from the direct association between the target ATU Topic and the ATU Numerical/ATU Type characteristics. Despite being included in the original dataset, these two variables represent categorical identifiers that are intrinsically connected to the categorization label. By include them directly, performance measurements would be artificially inflated because the model would have near-direct access to the solution. Consequently, ATU Type and ATU Numerical were not included as predictors during model training in our primary experimental runs. In order to ensure that the stated results accurately represent the model's capacity to learn from narrative and linguistic patterns rather than pre-encoded category information, the ATU Numerical and ATU Type were only kept for post-classification validation and interpretation.

Conclusion

Using a corpus of forty-four stories and a combination of machine learning methods, feature engineering, and text preparation, this study investigated the automatic categorization of Grimm's Fairy Tales. The textual data was prepared for analysis using preprocessing methods including transformation, tokenization, and filtering, and the Bag-of-Words model made it possible to express narrative content numerically. After evaluating many classification methods, such as k-Nearest Neighbors, Decision Tree, Random Forest, Naive Bayes, and Logistic Regression, the latter performed the best (AUC = 0.945, CA = 88%, F1 = 0.881).
By emphasizing the significance of certain words—such as "little," "fox," "cat," and "wolf"—to classification probabilities, the nomogram analysis provided interpretability and connected computational results with literary themes. While "fox" and "wolf" indicated recurrent animal archetypes with differing influences, "little" stood out as a powerful thematic signal associated with well-known heroes. This offers important insights on the lexical patterns that characterize the Grimm storytelling style in addition to validating the model's forecasting ability.
The results show that machine learning is capable of accurately and efficiently classifying literary works by modelling their structural and thematic components. Beyond scholarly study, these models may find use in automated archiving systems, digital humanities, and instructional programs that classify and suggest books according to thematic similarities. Limitations can be addressed in future work by combining bigger and more diverse corpora, experimenting with sophisticated embeddings (e.g., Word2Vec, BERT), integrating topic modelling to expand theme exploration, and decreasing possible bias from directly linked characteristics like the ATU type. Through the integration of quantitative machine learning techniques with qualitative literary interpretation, this study demonstrates how interdisciplinary methodologies may enhance our comprehension of cultural narratives and computational text analysis.

Works Cited

Buckingham, David. *Media Education: Literacy, Learning and Contemporary Culture*. Blackwell Publishers, 2004.

Byun, J. K.-S. "The Effects of Multimedia Fairy Tale and Narrative Fairy Tale Lectures on Children's Language Expression Ability and Drawing Representation Ability." *Journal of the Korea Academia-Industrial Cooperation Society*, vol. 15, no. 3, 2014.

Davis, Dan, et al. 'Follow the Successful Crowd: Raising MOOC Completion Rates through Social Comparison at Scale'. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, ACM, 2017, 454–63. *DOI.org (Crossref)*, https://doi.org/10.1145/3027385.3027411.

Doli-Kryeziu, Selvete. "Fairy Tales and the Lingual and Intellectual Development in Preschoolers." *European Journal of Language and Literature Studies*, vol. 9, no. 1, Jan.–June 2023.

Foster, Jane. *Computational Approaches to Folk Narratives*. Academic Press, 2022, 45.
Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.
Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing*. 3rd ed., Draft version, Stanford University, 2023, https://web.stanford.edu/~jurafsky/slp3/
Sebastiani, Fabrizio. "Machine Learning in Automated Text Categorization." *ACM Computing Surveys*, vol. 34, no. 1, 2002,1–47.
Tatar, Maria. *The Hard Facts of the Grimms' Fairy Tales*. Expanded 2nd ed., Princeton University Press, 2003.
Zipes, Jack. *The Brothers Grimm: From Enchanted Forests to the Modern World*. 2nd ed., Palgrave Macmillan, 2002.