Hits or Misses? A Linguistically Explainable Formula for Fanfiction Success

Giulio Leonardi[1,*,†], Dominique Brunato[2,†] and Felice Dell'Orletta[2,†]

[1]University of Pisa
[2]Istituto di Linguistica Computazionale "Antonio Zampolli", ItaliaNLP Lab, Pisa

Abstract

This study presents a computational analysis of Italian fanfiction, aiming to construct an interpretable model of successful writing within this emerging literary domain. Leveraging explicit features that capture both linguistic style and semantic content, we demonstrate the feasibility of automatically predicting successful writing in fanfiction and we identify a set of robust linguistic predictors that maintain their predictive power across diverse topics and time periods, offering insights into the universal aspects of engaging storytelling. This approach not only enhances our understanding of fanfiction as a genre but also offers potential applications in broader literary analysis and content creation.

Keywords

fanfiction, Italian corpus, success prediction, linguistic features, Explainable Boosting Machine

1. Introduction and Motivation

The growing proliferation of online literary content has led to the emergence of new genres and storytelling forms, with fanfiction being particularly popular among teens and young adults. Fanfiction consists of stories created by fans (mostly hobby authors) that extend or alter the narrative of existing popular media like books, movies, comics or games, and represents a significant portion of user-generated content on the web [1]. In recent years, the widespread popularity that this genre has assumed has prompted research into the linguistic and stylistic elements that contribute to its success, mirroring studies conducted on more traditional literary genres [2, 3, 4], *among others*.

Understanding the elements that contribute to narrative success is a fascinating area of research with implications across various fields, from literary analysis to digital humanities. From a socio-linguistic perspective, it can offer deeper insights into people and culture. It also has significant applications in areas such as personalized content recommendation and educational technology [5, 6]. While personal interests undoubtedly play a crucial role in predicting a reader's engagement with a literary content, the way information is presented can also evoke different reactions and levels of interaction, ultimately influencing the narrative's success. In this regards, recent advancements in Natural Language Processing (NLP) and machine learning offer a powerful lens for making explicit patterns that may explain the complex interplay between reader engagement and content success.

This paper moves in this field and presents a computational analysis focused on Italian fanfiction, addressing the following research questions: *i.)* Can the success of Italian fanfiction be automatically predicted using stylistic and lexical features of the texts?; *ii.)* Which types of features demonstrate the highest predictive capability, and how consistent are these features across different time periods and thematic domains?; *iii.)* To what extent can these features be explained in terms of their contribution to predicting success?

Our contributions. *i.)* We collected a corpus of Italian fanfiction stories enriched with metadata considered as proxies of their success; *ii.)* We investigate the relationship between stylistic and lexical features of stories and their success from a modeling perspective; *iii.)* We identified the most influential features in success prediction, showing the key role played by form and stylistic related features across time and thematic domains of fanfictions.

The paper is structured as follows: Section 2 briefly contextualizes our study among relevant literature; Section 3 presents the reference corpus of Italian fanfiction stories that we collected; in Section 4 we provide an overview of the approach we devised including the description of features used for classification and the classifiers employed. Section 5 discusses the main findings and offers a fine-grained analysis of the classification results in terms of feature explainability. In Section 6 we summarize key findings and outlining promising directions for future research in this field.

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
†These authors contributed equally.
g.leonardi5@studenti.unipi.it (G. Leonardi);
dominique.brunato@ilc.cnr.it (D. Brunato);
felice.dellorletta@ilc.cnr.it (F. Dell'Orletta)

2. Related Work

The exploration of online content and its engagement levels has increasingly benefited from advancements in NLP and machine learning. Different perspectives have been touched upon considering different textual domains, typology of linguistic features and quantitative metrics to operationalize a very subjective concept like success. The study by Toubia and colleagues [7] explores how the structure of narratives, particularly the internal semantic progression measured by features derived from dense word representations, affects the success of stories across different text typologies (movies, TV shows, and academic papers). Berger and colleagues [8] examine how the linguistic structure of online content affects user engagement, specifically by modeling sustainable attention. This concept goes beyond just attracting a reader with a catchy headline or advertisement; it also encompasses the likelihood that a reader will continue viewing or reading the content. In their analysis of more than 35,000 online contents from heterogeneous sources, they emphasize the role of features related to processing ease and emotional language.

In the realm of literary works, Ashok et al. [2] first leverage stylometric analysis and machine learning techniques to predict the success of popular English novels from the Gutenberg Project. Their approach demonstrated the potential of these techniques for assessing literary success. Extending these findings, Maharajan et al. [9] proposed a multi-task approach to simultaneously evaluating success and genre prediction. Using deep learning representations, in addition to hand-craft features related to topic, sentiment, writing style, and readability of books, they obtained better performance than the single success prediction task approach. Focusing on contemporary English-language literature, the study by Bizzoni and colleagues [10] investigate how perceived novel quality is influenced by a broad spectrum of textual features — such as those related to readability and sentiment — and how these perceptions vary depending on the reader's level of expertise.

The growing volume of online fanfiction has also been the subject of numerous studies, either from the perspective of text mining by using NLP or through a qualitative lens via a manual examination. A comprehensive survey of analyses in this direction has been recently provided by [11]. For example, Milli and Bamman [12] explore the relationship between fanfiction and its original canon, offering one of the first empirical analyses of this genre. Similarly, Sourati et al. [13] find that the similarity between fanfictions and their original stories — particularly in terms of emotional arcs and character dynamics—correlates significantly with fanfiction's popularity.

In the context of Italian fanfiction, research using NLP techniques is still limited. Mattei et al. [14] employ linguistic profiling to analyze a corpus of Italian fanfiction inspired by the Harry Potter series, with the purpose of identifying linguistic patterns associated with success. Inspired by this previous study, our research aims to extend these findings through a computational modeling approach, investigating the power of linguistic features for predicting fanfiction success and their generalization across different experimental settings.

3. Corpus Construction

As a first step, we compiled a reference corpus of Italian fanfiction. To this end, we searched available texts on efpfanfic.net, one of the largest Italian websites dedicated to publishing and reading amateur stories, focusing specifically on stories labeled in the fanfiction genre.

Using a web scraping system, we extracted fanfictions based on the *Harry Potter* series, a highly popular fandom on the site, boasting 57,196 stories published between 2003 and 2023. Figure 1 presents the temporal distribution of these fanfictions up to 2020.

Additionally, we gathered a secondary corpus consisting of 2,441 stories based on *The Lord of the Rings* series. This secondary corpus served as a test set to assess the influence of thematic domains on the analysis of story success.

For this study, we focused on the first chapter of each fanfiction to ensure a consistent analysis. While it is widely recognized that thematic units within stories — particularly the beginnings and endings — often differ from the middle sections due to their distinct narrative roles, we observed that the majority of stories (69%) consist of only a single chapter, making them effectively self-contained. The efpfanfic portal allows users to review each chapter with ratings marked as negative, neutral, or positive. Consistent with prior research such as [9] we used the absolute number of reviews to define the success of a story, which we consider broadly as popularity. This approach is based on the assumption that a high number of interactions, regardless of their sentiment, reflects strong reader's engagement. This is especially confirmed since in our dataset negative reviews represent less than 1% of the total.

To formulate our success prediction task, we established a review threshold to classify each story as either a success or a failure. After analyzing the distribution of reviews for *Harry Potter* texts (Figure 2), we decided to exclude stories that fell in the middle of the distribution – those that could not be clearly defined as successes or failures. Consequently, stories with fewer than two reviews (25th percentile) were classified as failures, and those with more than six reviews (75th percentile) as successes. Stories within the interquartile range were excluded from

Table 1
Descriptive Statistics for the Harry Potter (HP) and Lord of The Rings (LOTR) Corpora

Corpus	#texts	#negatives	#positives	avg. #tok
HP	26,032	13,058	12,974	1911
LOTR	932	526	406	1946

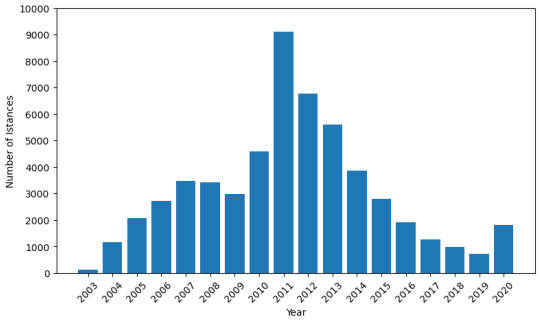

Figure 1: Distribution of all fanfictions from the Harry Potter corpus by year of publication (up to 2020).

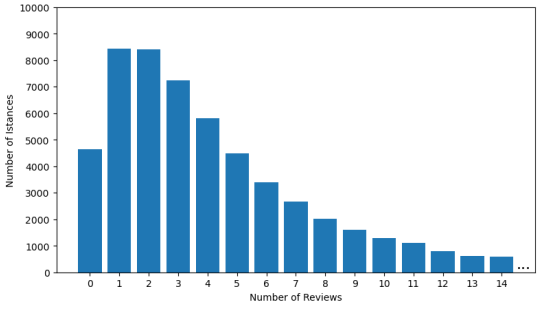

Figure 2: Distribution of published fanfiction from the Harry Potter corpus by number of reviews in the first chapter.

the analysis. We also excluded texts published after 2020, considering them too recent for meaningful comparison.

As summarized in Table 1, the final corpora, hereafter abbreviated as HP (Harry Potter) and LOTR (The Lord of the Rings), consist of 26,032 and 932 texts, respectively.

4. Methodology

Based on the newly collected dataset and its internal distinction, we formulated the task of success prediction as a binary classification problem, that is: given a story, the model is asked to predict whether it belongs to the successful or unsuccessful class, where the two classes were defined according to the metric based on the number of reviews received by readers.

In line with our main purpose to construct a model of success grounded on interpretable factors, we decided to leverage explicit features modelling both style-related and lexical aspects of text as input for the classification system. To evaluate the effectiveness and robustness of these features, we conducted experiments across three conceptually distinct scenarios to evaluate the ability to discriminate success in different contexts. Specifically, the first scenario is **in-domain**: the classifier is evaluated on texts within the same thematic domain as the training set, using 10-fold cross-validation on the HP corpus. The second scenario is **out-domain**: the classifier is evaluated on texts from a different thematic domain than the training set. In this case, the HP corpus is used as the training set, while the LOTR corpus serves as the test set.

Finally, in the **cross-time** scenario, the temporal impact on classification is considered. The classifier is trained solely on texts from the HP corpus published in 2011 and sequentially tested on texts from each other year from 2003 to 2020. The 2011 texts were chosen for training because this year has the largest amount of data (3,755 texts), is approximately central within the temporal range [2003, 2020], and is particularly significant for fanfiction production due to the release of the final film in the Harry Potter saga.

The main components of our approach are detailed in the following sections.

4.1. Success Predictors

A comprehensive set of features was extracted for each story in the corpus. These features were categorized into two primary groups: linguistic features, reflecting the text's linguistic style and structure and lexical features, representing the semantic content of the text.

4.1.1. Linguistic Features

To model text's linguistic style and structure, we drew inspiration from the linguistic profiling framework, a NLP-based methodology in which a large set of linguistically-motivated features automatically extracted from annotated texts is used to obtain a vector-based representation of it. Such representations can be then compared across texts representative of different textual genres and varieties to identify the peculiarities of each [15]. For our study, we relied on Profiling-UD[1], a multilingual tool inspired by this framework, which extracts over 130 linguistic features from texts using the Universal Dependencies (UD) annotation formalism. As described in Brunato et al. [16], these features encompass a range of linguistic phenomena that can be classified into distinct groups covering e.g. shallow text features (e.g. document and sentence length, average word length), distribution of grammatical categories, inflectional morphology and

[1] http://linguistic-profiling.italianlp.it/

syntactic properties related to local and global parse tree depth structure.

These features have proven effective in tasks related to modeling text form, such as assessing text complexity, and identifying stylistic traits of authors or author groups. Building on previous research on a similar corpus of fanfiction [14], we hypothesize that these features can also distinguish between successful and unsuccessful fanfictions from a modeling perspective.

4.1.2. Lexical Features

The second representation employed is based on lexical information and leverages the relative frequency of n-grams in each document. The choice of n-grams, in contrast to more powerful semantic representation derived from embeddings, is deliberately motivated by the desire to use lexical features that remain completely explicit. The model, henceforth referred to as the Lexical Model, consists of the following features:

- `Forms`: unigrams, bigrams, and trigrams of tokens.
- `Lemmas`: unigrams, bigrams, and trigrams of lemmas.
- `Characters`: sequences of characters at the beginning or end of words, ranging from 1 to 4 characters in length.

4.2. Classifiers

In line with our research questions, the explainability of the classification is crucial to evaluate the impact of linguistic and lexical features on the prediction of success. Therefore, two classification algorithms that allow for a precise global explanation of the predictions were selected.

The first classifier employed is a linear Support Vector Machine. By fitting a decision hyperplane in the feature space, this method enables the examination of the hyperplane's coefficients to assess the importance of the features.

The second algorithm employed is the Explainable Boosting Machine (EBM), which belongs to the family of Generalized Additive Models (GAMs). As explained in [17] a GAM is a model of the form:

$$g(y) = \beta_0 + \sum f_n(x_n) \tag{1}$$

where $g(.)$ is called the link function, used to model the output (e.g., the logistic function for classification). Each $f_n(.)$ is referred to as a shape function, which is a univariate function modeling the relationship between the feature n and the target.

The prediction is thus a sum of n non-linear and arbitrarily complex shape functions, generally resulting in better accuracy compared to linear models. Additionally, with a reasonable number of features, the model remains explainable. Each shape function can be visualized as a two-dimensional plot, with the feature value on the x-axis and the score assigned by the shape function on the y-axis. A score greater than 0 indicates a contribution towards the positive class, whereas a score less than 0 indicates a contribution towards the negative class. The final prediction value for a record is simply the sum of the scores obtained from each shape function, potentially transformed by the link function. Beyond analyzing individual shape functions, the average contribution of each feature can be evaluated by taking the mean of the absolute values of the assigned scores.

There are various algorithms within the family of GAMs, primarily distinguished by the method used to fit the shape functions. In the case of the EBM, standard gradient boosting is used. However, in each boosting iteration, the algorithm sequentially cycles through each feature, constructing each univariate shape function through bagged boosted trees. This method has proven to be one of the most effective for training a GAM.

For our study, the EBM was employed exclusively for experiments based on linguistic features due to the excessive dimensionality of the lexical model. This high dimensionality would have rendered the GAM too complex to interpret and too time-expensive to train.

Table 2
Classification Accuracy(%) of the Models. 'Ling.' and 'Lex.' refer respectively to models trained on linguistic and lexical features. The baseline corresponds to the majority class label.

Scenario	SVM Ling.	EBM Ling.	SVM Lex.	*Baseline*
in-domain	65.03	66.15	**69.95**	50.16
out-domain	59.22	**64.70**	43.45	56.43
avg. cross-time	62.02	**62.81**	49.31	49.20
average	62.09	**64.55**	54.24	51.93

5. Results and Discussion

The classification results are summarized in Table 2, for each model and scenario under evaluation.

For models using linguistic features, in the in-domain scenario both the SVM and the EBM outperform the majority class baseline, with accuracies of 65.03% and 66.15% respectively, compared to 50.16% for the baseline. This indicates that both classifiers are effectively capturing the linguistic patterns associated with success within the same thematic domain.

For linguistic models, in the out-domain scenario the performance of the SVM drops significantly, with an accuracy of 59.22%, whereas the EBM experiences a less

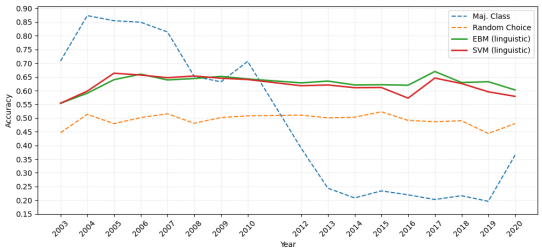

Figure 3: Classification Accuracy in the Cross-Time Setting

drastic decline, achieving an accuracy of 64.70%. However, both classifiers still perform better than the baseline, suggesting some degree of ability to generalize of the linguistic features across different thematic domains.

The lexical model, in the in-domain scenario, achieves an accuracy of 69.56%, outperforming all models with linguistic features, suggesting that lexical features provide a more powerful representation for in-domain success prediction. Nevertheless, in the out-domain scenario, the lexical model does not surpass the baseline, indicating a complete lack of predictive ability. This suggests that lexical features, which are primarily based on the content of the specific fanfiction's narrative universe, perform well within the same thematic domain but lose all significance outside of it. Conversely, linguistic features, which focus on the form of the text, appear to be more adaptable regardless of the theme.

Figure 3 presents the performance over time for classifiers trained with linguistic features. Additionally, two baselines are shown: "Random Choice", which randomly selects between the two classes, and "Maj. Class", which always assigns the majority class from the corresponding training set (2011 stories), i.e. the positive one. The results of the lexical model in the cross-time scenario were insignificant, as they were very similar to the "Maj. Class" baseline. The classifier, therefore, defaults to assigning the negative class, demonstrating no predictive capability. To avoid confusion, the lexical model results are not included in this Figure. In contrast, the cross-time results for models using linguistic features are more meaningful: the results remain stable around an average of 62%, regardless of the dominant class in the tested year and the classifier used (*avg. cross-time* in Table 2).

The cross-time scenario further suggests that linguistic features possess greater adaptability beyond their own domain, maintaining a considerable degree of generalization over time. Conversely, lexical features seem functional only within the specific domain of the training set, losing all predictive power for texts from different domains. Overall the model that performed best on average across the three scenarios, and with the least variance in performance, is the EBM trained with linguistic features. We provide an in-depth analysis of this model in the following section.

5.1. The Model of Success

To gain a better understanding of the classification results and identify the most influential features for predicting success, we ranked the features according to the absolute value of their weight in the EBM classifier model trained on the entire training set. Table 3 presents an extract of the top 15 features. The analysis reveals that, in addition to basic text features such as the average document length (measured in tokens [1]) and the average word length (in characters [2]), more complex linguistic properties play a crucial role. Among these, features related to verbal predicates and verbal morphology emerge as particularly influential. This suggests that the syntactic and morphological characteristics of verbs, such as tense, mood and person, provide valuable information for the classifier prediction, highlighting the importance of deeper linguistic structures in building a model of successful writing.

While this ranking highlights the 'global' importance of features, it does not explain their effect on classification. For a more detailed analysis, Figure 4 in Appendix A highlights the threshold values for each of the top 15 ranked features, indicating the point at which the expected classification shifts from one class to another. Additionally, it provides the number of instances in the training set for each feature value. Interestingly, there are some features which split almost exactly the amount of data into two subsets. For example, the features representing word length (*char_per_tok*) has a discriminant threshold of 4.55 characters which distinguishes successful stories – typically with longer words – from unsuccessful ones – usually with shorter words. Similarly, features related to the (morpho-)syntactic profile of the text such as the percentage of conjunctions (*dep_dist_conj*) and non-finite verb forms (*verbs_form_dist_Fin*) show a similar pattern. For these features, values lower than the discriminant threshold contribute to predicting the negative class, effectively splitting the data into two groups with comparable densities. Regarding verb presence (*verbal_head_per_sentence*), an increased use of verbs correlates with the unsuccessful class. This finding contradicts the idea that higher readability, typically conveyed by a predominantly verbal prose rather than a nominal one, is a good indicator of writing quality. However, it aligns with observations by Ashok et al. [2], who identified similar patterns in canonical literary novels.

Features related to verbal morphology also show a peculiar trend. For instance, a complementary perspective emerges concerning the use of person morphology. Increasing the use of second person plural beyond a relatively low threshold (0.4) positively affects the prediction

of success, which may indicate an alignment with the Reader-Insert[2] format, a specific type of fanfiction where the reader assumes the role of the protagonist, heavily relying on second-person narration. In contrast, an excessive use of the first person plural is associated with the negative class.

Table 3
Top 15 Scores of the EBM Trained with Linguistic Features

#	feature	score
#1	n_tokens	0.121
#2	char_per_tok	0.098
#3	verbal_root_perc	0.095
#4	verbs_num_pers_dist_Plur+2	0.090
#5	verbs_num_pers_dist_Plur+1	0.088
#6	upos_dist_SYM	0.080
#7	n_sentences	0.077
#8	aux_tense_dist_Imp	0.077
#9	verbs_tense_dist_Imp	0.072
#10	aux_tense_dist_Pres	0.067
#11	verbal_head_per_sent	0.066
#12	dep_dist_conj	0.065
#13	tokens_per_sent	0.064
#14	verbs_form_dist_Fin	0.053
#15	n_prepositional_chains	0.052

6. Conclusion

Understanding success factors in literary writing is an evolving area of cross-disciplinary research. This study on Italian fanfiction demonstrated the feasibility of predicting success using computational methods and explainability techniques. Notably, we found that features related to style and structure of texts show greater robustness than lexical ones across different domains and time periods. This suggests that the way a story is crafted may be more universally appealing than specific word choices or thematic elements.

We believe that the implications of this study extend far beyond fanfiction research. On the one hand, it provides new methodologies for analyzing online literary phenomena offering potential contributions to digital humanities. From the NLP perspective, it could inform text generation models, potentially guiding the creation of content that resonates more effectively with readers.

Future research could explore the generalizability of these findings to other languages and genres, as well as the investigation on the dynamics of evolving reader preferences over time by also considering alternative measures to gauge success. Additionally, this study does not take into account the importance of the author; a potential future development would be to consider the impact of the author's popularity and productivity on the success of their fanfiction.

[2] https://fanlore.org/wiki/Reader-Insert

References

[1] K. Hellekson, K. Busse, Fan fiction and fan communities in the age of the internet: new essays, McFarland, 2014.

[2] V. G. Ashok, S. Feng, Y. Choi, Success with style: Using writing style to predict the success of novels, in: Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1753–1764.

[3] J. Brottrager, A. Stahl, A. Arslan, U. Brandes, T. Weitin, Modeling and predicting literary reception. a data-rich approach to literary historical reception, Journal of Computational Literary Studies 1 (2022). URL: https://doi.org/10.48694/jcls.95.

[4] M. Algee-Hewitt, S. Allison, M. Gemma, R. Heuser, F. Moretti, H. Walser, Canon/archive : large-scale dynamics in the literary field, 2018. URL: https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf.

[5] Reviews matter: How distributed mentoring predicts lexical diversity on fanfiction.net, 2018. URL: https://api.semanticscholar.org/CorpusID:265096028.

[6] S. Sauro, Fan fiction and informal language learning, The handbook of informal language learning (2019) 139–151.

[7] O. Toubia, J. A. Berger, J. Eliashberg, How quantifying the shape of stories predicts their success, Proceedings of the National Academy of Sciences of the United States of America 118 (2021). URL: https://api.semanticscholar.org/CorpusID:235648521.

[8] J. A. Berger, W. W. Moe, D. A. Schweidel, What holds attention? linguistic drivers of engagement, Journal of Marketing 87 (2023) 793 – 809. URL: https://api.semanticscholar.org/CorpusID:255250393.

[9] S. Maharjan, J. Arevalo, M. Montes, F. A. González, T. Solorio, A multi-task approach to predict likability of books, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017, pp. 1217–1227.

[10] Y. Bizzoni, P. F. Moreira, I. M. S. Lassen, M. R. Thomsen, K. Nielbo, A matter of perspective: Building a multi-perspective annotated dataset for the study of literary quality, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 789–800. URL: https://aclanthology.org/2024.lrec-main.71.

[11] D. Nguyen, S. Zigmond, S. Glassco, B. Tran, P. J. Giabbanelli, Big data meets storytelling: using machine learning to predict popular fanfiction, Social Network Analysis and Mining 14 (2024) 58.

[12] S. Milli, D. Bamman, Beyond canonical texts: A computational analysis of fanfiction, in: J. Su, K. Duh, X. Carreras (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2048–2053. URL: https://aclanthology.org/D16-1218. doi:10.18653/v1/D16-1218.

[13] Z. Sourati Hassan Zadeh, N. Sabri, H. Chamani, B. Bahrak, Quantitative analysis of fanfictions' popularity, Social Network Analysis and Mining 12 (2022) 42.

[14] A. Mattei, D. Brunato, F. Dell'Orletta, The style of a successful story: a computational study on the fanfiction genre, in: J. Monti, F. Dell'Orletta, F. Tamburini (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, volume 2769 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2769/paper_52.pdf.

[15] H. van Halteren, Linguistic profiling for authorship recognition and verification, in: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain, 2004, pp. 199–206. URL: https://aclanthology.org/P04-1026. doi:10.3115/1218955.1218981.

[16] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, S. Montemagni, Profiling-UD: a tool for linguistic profiling of texts, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 7145–7151. URL: https://aclanthology.org/2020.lrec-1.883.

[17] Y. Lou, R. Caruana, J. Gehrke, Intelligible models for classification and regression, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2012). doi:10.1145/2339530.2339556.

A. Top 15 Features of the EBM

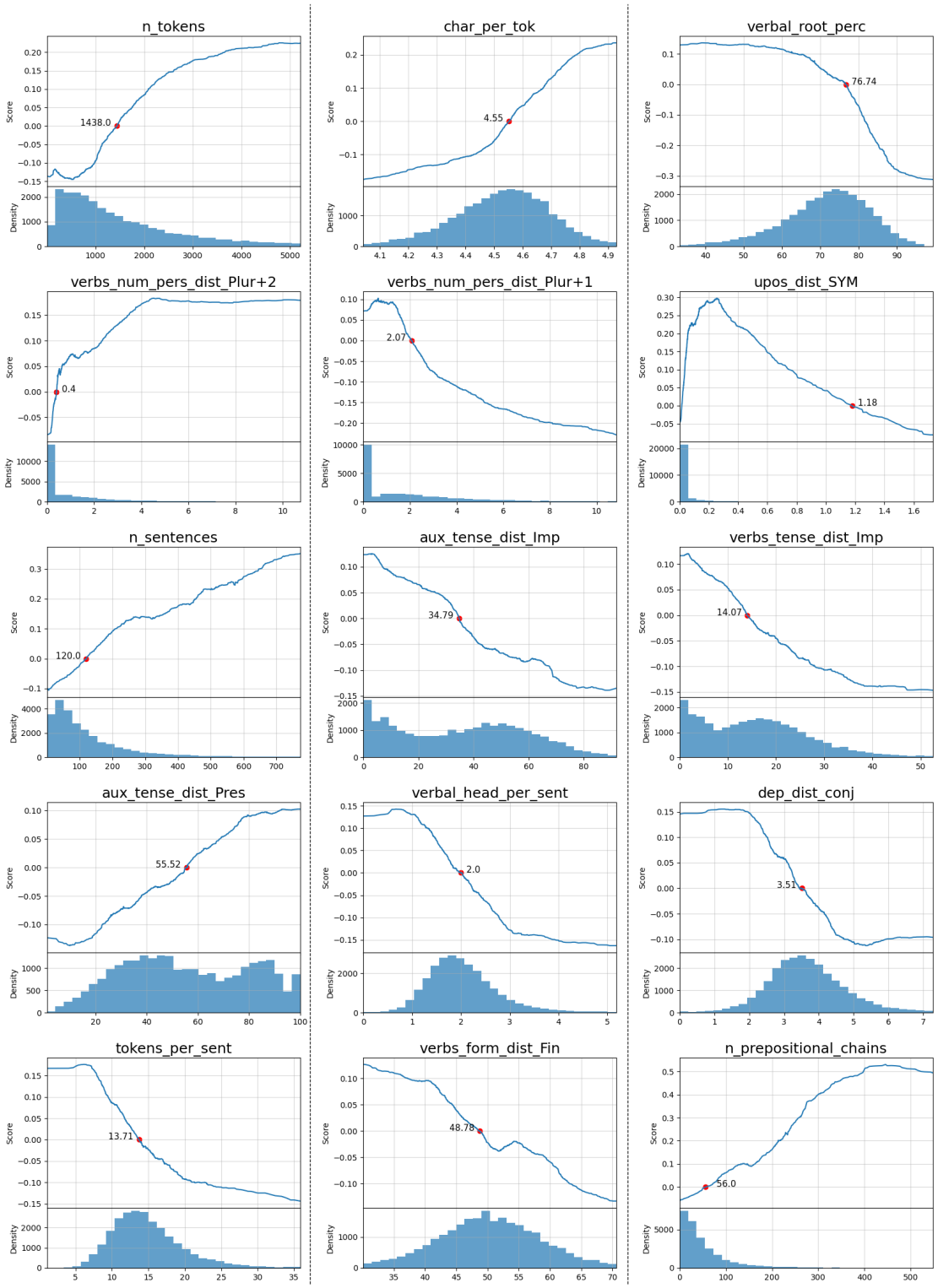

Figure 4: Visualization of the Shape Functions of the Top 15 Linguistic Features of the EBM. In each graph pair, the **x-axis** represents the feature value, the **y-axis** of the line plot indicates the score assigned by the shape function, and the marked threshold value denotes the feature value at the zero score point. For the features represented by absolute numbers (i.e. *n_tokens*, *char_per_tok*, *n_sentences*, and *n_prepositional_chains*), the values are displayed as raw counts. For the remaining features, which are expressed as percentage distributions, the values are shown accordingly. More details about how these features are calculated are reported in [16].