Classification of Text Writing Proficiency of L2 Learners

Giulio Biondi[1], Valentina Franzoni[1,2(✉)], Alfredo Milani[1], and Valentino Santucci[2,3]

[1] Department of Mathematics and Computer Science, University of Perugia, Via Vanvitelli 1, Perugia, Italy
{giulio.biondi,valentina.franzoni,alfredo.milani}@unipg.it
[2] Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China
valentino.santucci@unistrapg.it
[3] University for Foreigners of Perugia, Perugia, Piazza Grimana, Italy

Abstract. In this study, we present a novel system for the automatic classification of text complexity in the Italian language, focusing on the phraseological dimension. This quantitative assessment of text complexity is crucial for various applications, including text readability measurement, text simplification, and support for educators during evaluation processes. We use a dataset comprising texts written by Italian L2 learners and classified according to the levels of the Common European Framework of Reference for Languages. The dataset texts serve as a basis for calculating phraseological features, which are then used as input for multiple machine-learning classifiers to compare their performance in predicting proficiency levels. Our experimental results demonstrate that the proposed framework effectively harnesses phraseological complexity features to achieve high classification accuracy in determining proficiency levels.

Keywords: Text Complexity · Natural Language Processing · Text Classification

1 Introduction

In recent years, the analysis and evaluation of text complexity have garnered significant attention in the field of Natural Language Processing (NLP) and education, given its crucial role in various applications, such as readability assessment, text simplification, and assisting teachers in evaluation procedures [9,13,16,19]. The ability to accurately measure the complexity of a text is especially relevant for second language (L2) learners, as it allows for the selection of appropriate learning materials and the tailoring of teaching methods to individual student needs.

The Common European Framework of Reference for Languages (CEFR) [6] provides a widely recognized standard for language proficiency assessment,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14104, pp. 15–28, 2023.
https://doi.org/10.1007/978-3-031-37105-9_2

encompassing various skills such as reading, writing, listening, and speaking. However, the automatic classification of text complexity according to CEFR levels remains a challenging task. This challenge is further exacerbated when considering the phraseological dimension, a crucial aspect of language proficiency that has been relatively underexplored in previous research.

In this study, we aim to bridge this gap by proposing a novel system for automatically assessing text complexity in the Italian language, focusing specifically on the phraseological dimension and targeting CEFR proficiency levels. To achieve this goal, we utilize a dataset of texts produced by Italian L2 learners for written examinations, classified according to CEFR levels. We extract several phraseological features from the texts and employ multiple machine-learning classifiers to predict CEFR proficiency levels based on these features.

The rest of the paper is organized as follows: Sect. 2 provides an overview of related work in the field of text complexity assessment and phraseological analysis. Section 3 presents our methodology. Section 4 details the dataset, feature extraction, and machine learning classifiers used. Section 5 discusses the experimental results, demonstrating the effectiveness of the proposed framework in achieving high classification accuracy. Finally, Sect. 5 concludes the paper and outlines potential avenues for future research.

2 Related Work

In [19], an automatic classification system for measuring the complexity level of a given Italian text is introduced from a linguistic point of view. The task of measuring the complexity of a text is cast to a supervised classification problem by exploiting a dataset of textual instructions purposely produced by linguistic experts for second language teaching and assessment purposes. The commonly adopted levels from the Common European Framework of Reference for Languages were used as target classification classes, the exam instructions were elaborated by considering a large set of numeric linguistic features, and an experimental comparison among ten widely used machine learning models was conducted. The results show that the proposed approach is able to obtain good prediction accuracy on a dataset comprising lexical, morphological, syntactical, and morpho-syntactical features computed from a corpus of texts written by Italian L2 learners, while a further analysis was conducted in order to identify the categories of features that influenced the prediction. The maximum accuracy of 74.2% was achieved using a Random Forest classifier.

Previous attempts at classifying texts according to their complexity for the Italian language include [4,7]. The first presents a new resource for automatically assessing text difficulty in the context of Italian as a second or foreign language learning and teaching, called MALT-IT2, which automatically classifies input texts according to the CEFR level they are more likely to belong to. The second proposes a new approach to readability assessment with a specific view of the task of text simplification. The intended audience includes people with low literacy skills or mild cognitive impairment. READ-IT represents the first advanced

readability assessment tool for Italian, which combines traditional raw text features with lexical, morpho-syntactic, and syntactic information. In READ-IT, readability assessment is carried out for both documents and sentences where the latter represents an important novelty of the proposed approach creating the prerequisites for aligning the readability assessment step with the text simplification process.

Regarding measures of text complexity, in [16], the author proposes the use of Pointwise Mutual Information (PMI) [2] to determine the phraseological complexity of texts; PMI was also used Logdice association measure [18] estimating the complexity of reading comprehension texts classified with CEFR [14].

3 Method

In this work, the task of determining the complexity of a text is defined as a supervised classification problem; several machine learning classifiers are trained with features computed from texts extracted from the CELI corpus, [13] which includes texts of students' responses to the teachers' essay instructions. The texts in the corpus were produced by Italian L2 learners and are classified according to the learner's level, i.e. the CEFR proficiency levels B1, B2, C1, and C2. The texts are analyzed from the phraseological complexity point of view; more specifically, the work focuses on the text diversity and text sophistication dimensions.

3.1 Text Pre-processing

In Natural Language Processing applications, during the pre-processing phase techniques are employed to produce a structured output from unstructured texts; in particular, after tokenizing the texts, lemmatization and dependency parsing were performed to analyze the texts from the dimensions of interest. The texts in the CELI corpus were pre-processed by means of the UDPIPE software [20], coupled with the *italian-isdt-ud-2.10-220711* treebank. UDPIPE processes texts following the Universal Dependencies framework (UD) [15], which provides consistent grammar annotation rules for texts in several human languages; in particular, for texts in the Italian language, 17 Parts-of-speech tags and 45 dependency relations tags are defined. The following dependency relations were considered to analyze the phraseological complexity of the text:

- **obj**: the direct object of a verb
- **amod**: adjective modifying the meaning of a nominal head
- **advmod**: a (non-clausal) adverb or adverbial phrase modifying the meaning of the head.
- **conj**: a relation between two elements connected by a coordinating conjunction (e.g. *and* and *or*.

The conj dependency was considered to account for cases in which, for example, a verb had multiple direct objects, as in the following example taken from the dataset:

Cambiare stile di vita, luogo, lavoro è sempre comunque un'esperienza positiva che ci rendi più forti.
Changing lifestyle, place, job is, in any case, a positive experience that strengthens us.
As visible in Fig. 1, the word *stile* is tagged by the parser as a direct object (*obj*)

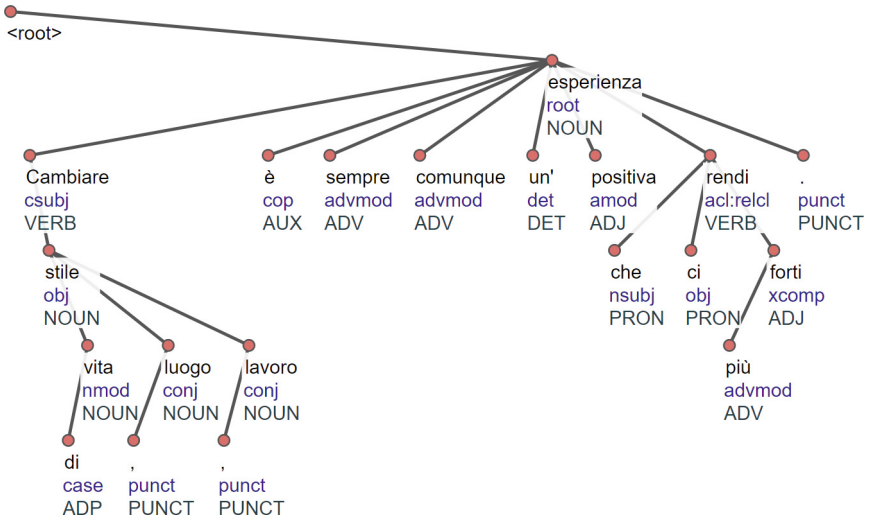

Fig. 1. UDPIPE parsing tree of the example sentence

of the head verb *cambiare*, while *luogo* and *lavoro* as coordinated (*conj*) with the head noun *stile*, although both are syntactically direct objects of *cambiare* too. For this reason, conjuncts with words tagged as *obj*, *amod*, or *advmod* were also considered for the computation of statistics. Words selected according to the criteria mentioned above were selected and used to compute the features in Sects. 3.2 and 3.3.

3.2 Text Sophistication

According to [16], the text sophistication dimension can be calculated [1,17] following two main approaches: the first is based on selecting uncommon word combinations in a text according to a list of sophisticated word collocations or by exclusion from a list of most frequent combinations; however, such resources are usually not available for a language, including Italian. Another strategy requires estimating the association between words by means of an appropriate frequency-based word association measure; collocations scoring high association values usually carry a rather distinctive meaning, and can therefore be associated with the sophistication level of a text. In this work, a frequentist approach is adopted, and the sophistication of collocations is estimated through the

Pointwise Mutual Information (PMI) or LogDice measures. Pointwise Mutual Information is an association measure frequently employed in NLP applications; it measures the association between two terms x and y with their occurrence and co-occurrence values as measured in a sufficiently large reference corpus; it is defined as

$$PMI = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

where the numerator is the co-occurrence probability and the denominator the product of the occurrence probabilities of the two terms. In this work, two tokens x and y, where $xDEPy$ for some dependency DEP, are considered co-occurring each time they appear under the same dependency link in the reference corpus. PMI has been used in literature for several NLP applications, including text emotion recognition [8], topic modelling [22] and text generation [21]. LogDice is a standardized measure to quantify the tendency of two tokens to occur together in a corpus, defined as:

$$14 + \log_2 \frac{2 * O_{11}}{R1 + C1}$$

where O_{11} is the observed frequency of the collocation and $R1$ and $C1$ are the frequencies of the two tokens in the reference corpus. In this work the Paisa corpus [11] was used as the reference corpus, to draw the occurrence and co-occurrence values used to compute PMI/Logdice association measures for each pair of words.

3.3 Text Diversity

The phraseological diversity dimension can be substantially measured in analogy with lexical diversity, by defining adequate type-token ratio measures. To measure the lexical diversity of a text, Root Type Token Ratio (RTTR) is frequently used in NLP to measure the degree of lexical variation and is defined as:

$$RTTR = \frac{t}{\sqrt{n}}$$

where t is the count of different types in the text and n is the total number of tokens. In this work the following variations of RTTR were also considered and employed as features:

- Lemma-TTR, where t is the count of different lemmas in the text;
- Obj-TTR, where t is the count of different tokens dependency-tagged as *obj* and n the total count of tokens tagged as *obj*;
- Amod-TTR, where t is the count of different tokens dependency-tagged as *amod* and n the total count of tokens tagged as *amod*;
- Advmod-TTR, where t is the count of different tokens dependency-tagged as *advmod* and n the total count of tokens tagged as *advmod*;

4 Experiments Settings

4.1 Dataset

The dataset is composed of 3041 texts, with an average of 12.3913 sentences and ≈ 210 tokens per text. 13 texts contain no occurrences of one or more of the considered dependencies, and thus the calculation of the corresponding measures was impossible. For this reason, such texts were excluded, bringing the total down to 3028.

In Table 1 the class distribution of the dataset is shown; the dataset is unbalanced, with B1 accounting for more than 33% of the total texts, and C2, i.e. the highest proficiency level being the less represented class. In Table 2 the number

Table 1. Dataset statistics

Class	#Texts	#Types	#Tokens
B1	1212	6656	161776
B2	840	8484	159017
C1	585	11377	158289
C2	404	12782	159120
Total	3041	22039	638202

of tokens tagged with each of the dependencies and the total number of tokens in the CELI Corpus are reported; ≈ 16% of the tokens were selected for further analysis [3,5].

Table 2. Tokens by dependency

Dependency	Count
Obj	35747
Amod	27714
Advmod	42473
Conj/Obj	1722
Conj/Amod	1214
Conj/Advmod	77
Total	638202

4.2 Features

Five different features combinations were extracted, to assess the contribution of the sophistication and diversity aspects in determining the complexity of a text:

- **Test 1**: diversity features only
- **Test 2**: Sophistication (PMI) features
- **Test 3**: Sophistication (PMI) + diversity features
- **Test 4**: Sophistication (LogDice) features
- **Test 5**: Sophistication (LogDice) + diversity features

4.3 Classification Models

In this work, for each feature subset, 12 different classification models, listed in Table 3, were trained: Support Vector Machines classify records by projecting

Table 3. Classification models

Model	Description
SVM_Linear	Support Vector Machine with Linear kernel
SVM_RBF	Support Vector Machine with Radial Basis Function kernel
SVM_Poly	Support Vector Machine with Polynomial kernel
RF_Gini	Random Forest classifier with Gini criterion
RF_Entropy	Random Forest classifier with Entropy criterion
RF_Log_loss	Random Forest classifier with Log_loss criterion
MLP_1	Multi Layer Perceptron with one hidden layer
MLP_2	Multi Layer Perceptron with two hidden layers
MLP_3	Multi Layer Perceptron with three hidden layers
NB	Naive Bayes classifier fitting Gaussian distributions
QDA	Quadratic Discriminant Analysis classifier
AdaBoost	AdaBoost boosting classifier

them in a higher dimensional space where they are separable by a set of hyperplanes. The choice of the kernel function determines how the original records are represented in the vector space; in this work, three different kernels, i.e. Linear, Radial Basis Function, and Polynomial were tested. Random Forest is an ensemble classification method that fits a large set of decision trees (100 in this work) on a dataset and adopts a majority classification scheme to aggregate the outputs of the decision trees and classify an unknown record. Decision trees are built by learning decision rules which iteratively split data into different subsets. During the tree learning process, the quality of each split is measured by a criterion; the Gini, Entropy, and Log_loss criteria were considered here. The Multi-Layer Perceptron (MLP) classifier is a conventional feed-forward neural network trained with the gradient descent algorithm. The input layer has as many neurons as the number of features fed to the classifier, and the output layer has one neuron per different class (in this work four, corresponding to the CEFR B1 to C2 classes); between the input and the output layer other hidden layers are responsible for learning patterns in the data. Three models, with a different number of hidden layers, were designed in accordance with experiments

in [19]; for MLP_1 one hidden layer of 25 neurons was used, while MLP_2 and MLP_3 introduce respectively one and two additional hidden layers of the same size. The AdaBoost Classifier is an ensemble method fitting a series of weak estimators on the same dataset, focusing iteratively on more difficult and specific records; in its default Scikit-Learn settings, the base estimator for Adaboost is a Decision Tree classifier with maximum depth 1. Finally, Gaussian Naive Bayes and QDA are two classification algorithms based on Bayesian theory; the former aims at learning a Bayesian network model to perform predictions, while the latter builds, for each of the target classes, a multivariate probabilistic model and classifies unknown records by maximizing the posterior probability.

4.4 Performance Metrics

The performance of the model is evaluated through the common metrics used for classification tasks, i.e. Accuracy, Precision, Recall, and F1-Score. The confusion matrices obtained by the best classifier for each feature subset are also reported. The scores are calculated by adopting a 10-fold Cross-Validation approach: each text belongs to one test set and is classified using the model trained on the corresponding training set.

5 Experimental Results

In Table 4 the results of the experiments considering diversity features only are reported; QDA proves to be the best performer according to all the metrics. There are no significant differences among classifiers, except for AdaBoost and Gaussian Naive Bayes, showing sensibly lower figures.

Table 4. Test 1 (diversity) results

Classifier	Accuracy	Precision	Recall	F1
SVM, Linear	0.6889	0.6839	0.6709	0.6756
SVM, RBF	0.6869	0.6907	0.6600	0.6691
SVM, Poly	0.6889	0.6858	0.6688	0.6743
RF, Gini	0.6711	0.6664	0.6528	0.6578
RF, Entropy	0.6622	0.6563	0.6424	0.6476
RF, Log_loss	0.6622	0.6563	0.6424	0.6476
MLP_1	0.6849	0.6814	0.6630	0.6688
MLP_2	0.6846	0.6788	0.6630	0.6677
MLP_3	0.6879	0.6851	0.6674	0.6708
AdaBoost	0.6354	0.6065	0.6088	0.5952
Gaussian Naive Bayes	0.6539	0.6408	0.6381	0.6380
QDA	**0.6922**	**0.6886**	**0.6744**	**0.6793**

Table 5. Test 1 normalized confusion matrix, QDA (diversity)

QDA	B1	B2	C1	C2
B1	**0.8421**	0.1397	0.0175	0.0008
B2	0.3792	**0.5024**	0.1100	0.0084
C1	0.0530	0.1795	**0.6427**	0.1248
C2	0.0099	0.0248	0.2550	**0.7104**

Table 6. Test 2 (sophistication-PMI) results

Classifier	Accuracy	Precision	Recall	F1
SVM, Linear	0.7332	0.7303	0.7200	0.7246
SVM, RBF	0.7325	0.7318	0.7169	0.7234
SVM, Poly	**0.7381**	**0.7377**	0.7164	**0.7247**
RF, Gini	0.7249	0.7229	0.7120	0.7165
RF, Entropy	0.7242	0.7212	0.7105	0.7148
RF, Log_loss	0.7242	0.7212	0.7105	0.7148
MLP_1	0.7351	0.7273	0.7196	0.7229
MLP_2	0.7358	0.7346	**0.7212**	0.7272
MLP_3	0.7318	0.7252	0.7157	0.7196
AdaBoost	0.6625	0.6419	0.6442	0.6391
Gaussian Naive Bayes	0.7157	0.7103	0.6964	0.7019
QDA	0.7199	0.7161	0.7015	0.7070

Table 7. Test 2 normalized confusion matrix, SVM Poly (PMI)

SVM_Poly	B1	B2	C1	C2
B1	**0.8969**	0.0956	0.0075	0.0000
B2	0.3325	**0.5634**	0.0981	0.0060
C1	0.0308	0.2325	**0.6479**	0.0889
C2	0.0025	0.0223	0.2178	**0.7574**

In Table 6 the results of the experiments including the text sophistication features are reported; the confusion matrix for the best classifier, i.e. SVM with linear kernel, is reported in Table 7. In Table 8 the results of the experiments including the base, text sophistication, and text diversity features are reported (Table 5). The confusion matrix for the best-performing classifier, i.e. SVM with Linear Kernel, is shown in Table 9. In Fig. 2 the F1 scores of all the classifiers are plotted for Test 5, i.e. LogDice-based sophistication and diversity features; the best-performing group is represented by SVM classifiers, followed by Random Forests, MLP3 and QDA groups with substantially similar performance, while

Gaussian Naive Bayes and, most notably, AdaBoost generally achieve lower figures. Such behaviour is generally consistent across experiments, with notable exceptions being MLP1 and MLP2 scoring higher than some RF-based classifiers in Test 1 and Test 2. Complete results for Test 5 are reported in Table 12; the best performance is achieved by SVM with linear kernel, with an Accuracy of 76.85%, 2.7% higher than the best one so far for Italian CEFR reported in [19]. Furthermore, in linguistics the CEFR B1, B2 and C1, C2 levels are often grouped in two macro-levels B and C; the confusion matrix for Test 5 in Table 13 shows that misclassification errors occur mostly between classes of the same macro-level and distant classes, e.g. B1 and C2, are seldom confused. If the classification output is aggregated according to the grouped categorization, the system achieves an overall accuracy of $\approx 93\%$; a similar behaviour can be observed for all the test cases (Tables 10 and 11).

Table 8. Test 3 (diversity+sophistication-PMI) results

Classifier	Accuracy	Precision	Recall	F1
SVM, Linear	0.7589	0.7597	**0.7508**	**0.7545**
SVM, RBF	0.7517	0.7544	0.7378	0.7444
SVM, Poly	**0.7606**	**0.7639**	0.7468	0.7534
RF, Gini	0.7325	0.7322	0.7213	0.7261
RF, Entropy	0.7312	0.7306	0.7158	0.7220
RF, Log_loss	0.7312	0.7306	0.7158	0.7220
AdaBoost	0.6628	0.6393	0.6416	0.6320
Gaussian Naive Bayes	0.7223	0.7132	0.7128	0.7121
QDA	0.7378	0.7338	0.7212	0.7259
MLP_1	0.7272	0.7174	0.7092	0.7127
MLP_2	0.7348	0.7277	0.7189	0.7230
MLP_3	0.7437	0.7417	0.7330	0.7366

Table 9. Test 3 normalized confusion matrix, SVM Linear (diversity+PMI)

SVM_Linear	B1	B2	C1	C2
B1	**0.8587**	0.1313	0.0100	0.0000
B2	0.2763	**0.6148**	0.1005	0.0084
C1	0.0239	0.1538	**0.7350**	0.0872
C2	0.0025	0.0198	0.1832	**0.7946**

Table 10. Test 4 (sophistication-LogDice) results

Classifier	Accuracy	Precision	Recall	F1
SVM, Linear	0.7368	0.7316	**0.7215**	0.7261
SVM, RBF	0.7361	0.7356	0.7187	0.7262
SVM, Poly	**0.7424**	**0.7410**	0.7185	**0.7278**
RF, Gini	0.7269	0.7185	0.7074	0.7123
RF, Entropy	0.7325	0.7254	0.7144	0.7193
RF, Log_loss	0.7325	0.7254	0.7144	0.7193
MLP_1	0.7338	0.7239	0.7158	0.7192
MLP_2	0.7341	0.7260	0.7165	0.7207
MLP_3	0.7384	0.7317	0.7219	0.7262
AdaBoost	0.6648	0.6328	0.6396	0.6296
Gaussian Naive Bayes	0.7249	0.7160	0.7032	0.7084
QDA	0.7269	0.7205	0.7077	0.7128

Table 11. Test 4 normalized confusion matrix, SVM Poly (LogDice)

SVM_Poly	B1	B2	C1	C2
B1	**0.8978**	0.0948	0.0075	0.0000
B2	0.3086	**0.5933**	0.0909	0.0072
C1	0.0308	0.2564	**0.6256**	0.0872
C2	0.0074	0.0198	0.2153	**0.7574**

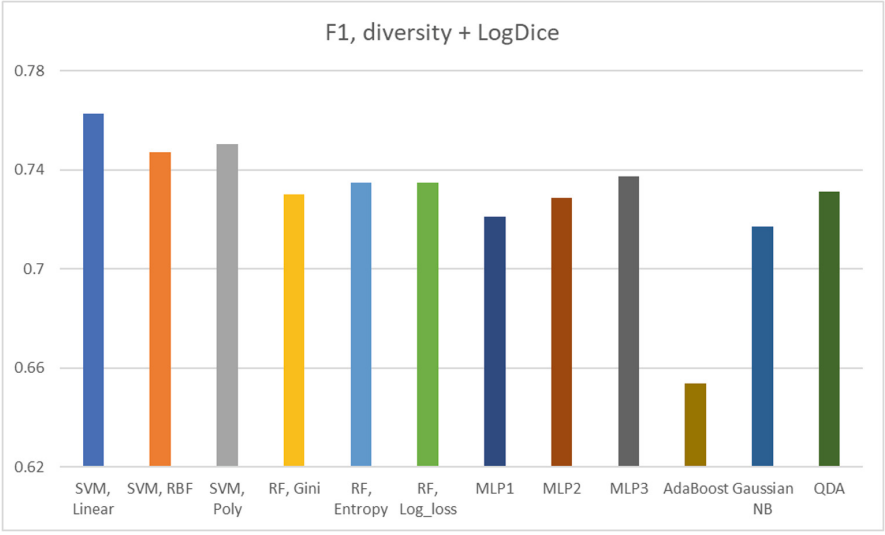

Fig. 2. Classifiers' F1 Score for Test 5, (features: diversity + sophistication-LogDice)

Table 12. Test 5 (diversity + sophistication-LogDice) results

Classifier	Accuracy	Precision	Recall	F1
SVM, Linear	**0.7685**	**0.7684**	**0.7583**	**0.7628**
SVM, RBF	0.7550	0.7562	0.7406	0.7470
SVM, Poly	0.7599	0.7619	0.7428	0.7503
RF, Gini	0.7394	0.7372	0.7243	0.7301
RF, Entropy	0.7447	0.7433	0.7287	0.7349
RF, Log_loss	0.7447	0.7433	0.7287	0.7349
MLP_1	0.7375	0.7269	0.7165	0.7211
MLP_2	0.7427	0.7377	0.7224	0.7288
MLP_3	0.7470	0.7405	0.7352	0.7372
AdaBoost	0.6800	0.6593	0.6583	0.6537
Gaussian Naive Bayes	0.7289	0.7185	0.7173	0.7170
QDA	0.7444	0.7382	0.7277	0.7313

Table 13. Test 5 normalized confusion matrix, SVM Linear (diversity+LogDice)

SVM_Linear	B1	B2	C1	C2
B1	**0.8703**	0.1205	0.0091	0.0000
B2	0.2596	**0.6352**	0.0969	0.0084
C1	0.0205	0.1675	**0.7282**	0.0838
C2	0.0025	0.0223	0.1757	**0.7995**

6 Conclusions

In this work, a framework is proposed for the automated classification of texts produced by Italian language L2 learners according to the Common European Framework of Reference (CEFR) B1 to C2 levels. The proposed approach focuses on phraseological text complexity features. Phraseological complexity is defined by the notions of text sophistication and text diversity, which are modelled with appropriate statistical metrics (i.e. Pointwise Mutual Information, LogDice, Type-token Ratios), applied to lemmas, objects, adjectives, and adverbial modifiers in the text. The effectiveness of this approach and possible metrics and feature sets has been experimentally evaluated on various state-of-the-art classifier algorithms. The CELI corpus, which is a dataset of real texts written by L2 students in official CEFR exams and classified with the candidates' levels of Italian competence certification, has been used in experiments.

Results show a promising capability of the framework to assess the text complexity level with different sets of features; the best outcomes were obtained by combining diversity features and sophistication features based on LogDice measures, with an accuracy of 76.85% and an F1 of 76.28% when using Support

Vector Machines. For CEFR Italian language L2 learners, the proposed strategy improves previous results obtained in the literature [19].

Future work will explore whether it is possible to further improve performance, with the integration of the proposed pure phraseological framework with an approach considering lexical and morphological quantitative indices as that in [19]; another potentially promising innovative direction is to explore the role of text embedding vectorial representations [10,12] in reflecting text complexity.

Acknowledgment. This work is partially supported by the Italian Ministry of Research under PRIN Project "PHRAME" Grant n.20178XXKFY.

References

1. Carpi, A., D'Alessandro, F.: On the hybrid Černý-road coloring problem and Hamiltonian paths. In: Gao, Y., Lu, H., Seki, S., Yu, S. (eds.) DLT 2010. LNCS, vol. 6224, pp. 124–135. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14455-4_13
2. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguist. **16**(1), 22–29 (1990)
3. D'Alessandro, F., Carpi, A.: On incomplete and synchronizing finite sets. Theoret. Comput. Sci. **664**, 67–77 (2017). https://doi.org/10.1016/j.tcs.2015.08.042
4. Dell'Orletta, F., Montemagni, S., Venturi, G.: READ-IT: assessing readability of Italian texts with a view to text simplification. In: Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, pp. 73–83. Association for Computational Linguistics, Edinburgh, Scotland, UK (2011)
5. Carpi, A., D'Alessandro, F.: On the commutative equivalence of context-free languages. In: Hoshi, M., Seki, S. (eds.) DLT 2018. LNCS, vol. 11088, pp. 169–181. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98654-8_14
6. Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division: Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press, Cambridge (2001)
7. Forti, L., Grego Bolli, G., Santarelli, F., Santucci, V., Spina, S.: MALT-IT2: a new resource to measure text difficulty in light of CEFR levels for Italian L2 learning. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 7204–7211. European Language Resources Association, Marseille, France (2020)
8. Franzoni, V., Milani, A., Biondi, G.: SEMO: a semantic model for emotion recognition in web objects. In: Proceedings of the International Conference on Web Intelligence, pp. 953–958. WI 2017, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3106426.3109417
9. Franzoni, V., Poggioni, V., Zollo, F.: Automated classification of book blurbs according to the emotional tags of the social network zazie. ESSEM@ AI* IA **1096**, 83–94 (2013). CEUR-WS
10. Li, Y., Yang, T.: Word embedding for understanding natural language: a survey. In: Srinivasan, S. (ed.) Guide to Big Data Applications. SBD, vol. 26, pp. 83–104. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-53817-4_4
11. Lyding, V., et al.: The PAISÀ corpus of Italian web texts. In: Proceedings of the 9th Web as Corpus Workshop (WaC-9), pp. 36–43. Association for Computational Linguistics, Gothenburg, Sweden (2014). https://doi.org/10.3115/v1/W14-0406

12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, vol. 2, pp. 3111–3119. NIPS 2013, Curran Associates Inc., Red Hook, NY, USA (2013)
13. Milani, A., Franzoni, V., Biondi, G.: Parsing tools for Italian phraseological units. In: Gervasi, O., et al. (eds.) ICCSA 2021. LNCS, vol. 12955, pp. 427–435. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87007-2_30
14. Natova, I.: Estimating CEFR reading comprehension text complexity. Lang. Learn. J. **49**(6), 699–710 (2021). https://doi.org/10.1080/09571736.2019.1665088
15. Nivre, J., et al.: Universal dependencies v2: an evergrowing multilingual treebank collection. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 4034–4043. European Language Resources Association, Marseille, France (2020)
16. Paquot, M.: The phraseological dimension in interlanguage complexity research. Second. Lang. Res. **35**(1), 121–145 (2019). https://doi.org/10.1177/0267658317694221
17. Poggioni, V., Bartoccini, U., Carpi, A., Santucci, V.: Memes evolution in a memetic variant of particle swarm optimization. Mathematics **7**(5), 423 (2019). https://doi.org/10.3390/math7050423
18. Rychlý, P.: A lexicographer-friendly association score. In: RASLAN (2008)
19. Santucci, V., Santarelli, F., Forti, L., Spina, S.: Automatic classification of text complexity. Appl. Sci. **10**(20), 7285 (2020). https://doi.org/10.3390/app10207285
20. Straka, M.: UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 197–207. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/K18-2020
21. Takayama, J., Arase, Y.: Relevant and informative response generation using pointwise mutual information. In: Proceedings of the First Workshop on NLP for Conversational AI, pp. 133–138. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/W19-4115
22. Vishal, K., Deepak, G., Santhanavijayan, A.: An approach for retrieval of text documents by hybridizing structural topic modeling and pointwise mutual information. In: Mekhilef, S., Favorskaya, M., Pandey, R.K., Shaw, R.N. (eds.) Innovations in Electrical and Electronic Engineering. LNEE, vol. 756, pp. 969–977. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-0749-3_74