


Article

Classification of Literary Works: Fractality and Complexity of the Narrative, Essay, and Research Article

Aldo Ramirez-Arellano 

Sección de Estudios de Posgrado e Investigación, Unidad Profesional Interdisciplinaria de Ingeniería y Ciencias Sociales y Administrativas, Instituto Politécnico Nacional, Ciudad de México 07738, Mexico; aramirez@ipn.mx; Tel.: +52-5557296000

Received: 22 July 2020; Accepted: 15 August 2020; Published: 17 August 2020



Abstract: A complex network as an abstraction of a language system has attracted much attention during the last decade. Linguistic typological research using quantitative measures is a current research topic based on the complex network approach. This research aims at showing the node degree, betweenness, shortest path length, clustering coefficient, and nearest neighbourhoods' degree, as well as more complex measures such as: the fractal dimension, the complexity of a given network, the Area Under Box-covering, and the Area Under the Robustness Curve. The literary works of Mexican writers were classified according to their genre. Precisely 87% of the full word co-occurrence networks were classified as a fractal. Also, empirical evidence is presented that supports the conjecture that lemmatisation of the original text is a renormalisation process of the networks that preserve their fractal property and reveal stylistic attributes by genre.

Keywords: complex networks; literary works; genre classification; stylistic attributes; lemmatization; renormalisation process

1. Introduction

A complex network as an abstraction of a language system has attracted attention in the last decade. The current linguistics research, based on the complex network approach, follows three major lines [1,2]: characterisation of human language as a multi-level system, linguistic typological research using quantitative measures, and the relationship between system-level complexity of human language and its microscopic features.

Word co-occurrence networks and their measures have been widely employed to analyse the syntactic features for multiple purposes, such as: identifying authors' writing styles [3–8] and evaluating machine translations [9]. Also, Ferraz de Arruda, Nascimento Silva [10], as well as F. de Arruda, Q. Marinho [11] built a complex network where the nodes are the representation of adjacent paragraphs that share a minimum semantical content to classify the text as real (written by an author) or randomly constructed (built from random blocks of real texts).

In most of the research mentioned above, well-known measures such as: node degree (k), shortest path length (spl), betweenness (b), clustering coefficient (cc), and the average of nearest neighbourhoods' degree (nnd) are applied to characterise the word co-occurrence networks. The k , b , and nnd are centrality measures that characterise local properties of the network that are useful for authorship attribution [3–8]. However, these measures do not capture the global network structure that could give us insight into the literary genre. This research aims at showing that local and global measures of the word co-occurrence networks—of literary works of Mexican writers—let us classify them according to the genre. Thus, the following research questions are formulated:

1. Are measures of the complex network useful to classify literary works by genre?
2. Is the full word co-occurrence network of literary works fractal?
3. Do pre-process tasks such as: deletion of number, punctuation, functional words, and lemmatisation generate fractal networks?

2. Measures of Complex Networks

Formally, a network is defined by $G = (V, E)$ where V is the vertexes or nodes, and E is the edges. The complex networks exhibit non-trivial topological features that do not occur in simple networks, such as: lattices or random graphs [12], and their overall behaviour cannot be predicted by observing the behaviour of their nodes [13]. Since the complex network theory has its root in graph theory, some measures are presented below.

The degree of a node i is defined by:

$$k_i = \sum_j^N v_{ij} \quad (1)$$

where j represents a given neighbour of the node i , and N is the total neighbours. The value of v_{ij} is defined as one, if there is a connection between nodes i and j , and as 0 otherwise.

Similarly, the betweenness of a node is defined as:

$$b_i = \sum_{j,m \neq i} \frac{L_{jm}(i)}{L_{jm}} \quad (2)$$

where L_{jm} is the number of shortest paths between nodes j and k , and $L_{jm}(i)$ is the shortest paths between nodes j and m that go through i .

The average nearest neighbourhoods' degree (nnd) of a given node can be computed by:

$$nnd_i = \sum_{j \in V(i)} \frac{k_j}{k_i} \quad (3)$$

where k_i is the degree of the node i , and the set $V(i)$ contains its nearest neighbours, and k_j is the degree of a given neighbour.

A definition of network clustering is expressed by:

$$cc(G) = \frac{3\tau}{spl(2)} \quad (4)$$

where τ is the number of triangles of the network and $spl(2)$ is the shortest path of length two. A "triangle" is a set of three nodes in which each contacts the other two.

2.1. Fractality of Complex Networks

A fractal is an object that is similar to itself on all scales [14]. A network is a fractal network if its box-covering follows the power law given by:

$$N_b(l) \sim \beta l^{-d_b} \quad (5)$$

where $N_b(l)$ is the minimum number of boxes of diameter l to cover the network—the procedure of box-covering that gives us this number is detailed later— β is the scaling factor, and d_b is the box dimension of a complex network that can be obtained as follows:

$$d_b = -\lim_{l \rightarrow 0} \frac{\ln N_b(l)}{\ln l} \tag{6}$$

On the other hand, a non-fractal network is characterised by a sharp decay of $N_b(l)$, with l described by an exponential function as follows [15,16]:

$$N_b(l) \sim \beta e^{-d_b l} \tag{7}$$

2.2. Complexity of Networks

The complexity measure of a network proposed by Lei, Liu [17] is defined as:

$$c(G) = d(G)s(G) \tag{8}$$

where $d(G) = |E| / (4CR^3 / 3\Delta)$ is the absolute density [18]; $|E|$, C , R , and Δ are the number of edges, circumference, radius, and diameter of the network, respectively. $s(G) = -k \sum_{i=1}^{|V|} (p_i^{q_i} - p_i) / (1 - q_i)$ is known as structure entropy based on degree and betweenness [17], where k is the Boltzmann constant, $|V|$ is the number of nodes, $p_i = \frac{k_i}{\sum_{i=1}^{|V|} k_i}$, $q_i = 1 + (b_{max} - b_i)$, and b_{max} is the maximum value of the betweenness computed by the Equation (2). This measure captures the topology of the networks, but it is not affected by scales and their types.

2.3. Box-Covering of Complex Networks

To obtain $N_b(l)$, consider the phrase “No one behind, no one ahead”. Its word co-occurrence network is shown in Figure 1. The number of boxes to cover the network $N_b(l)$ for $l = 1$, and $l = \Delta + 1$ —where Δ is the diameter of the network—is the number of nodes of the network and one, respectively. The $N_b(l)$ from 2 to Δ is not a trivial answer.

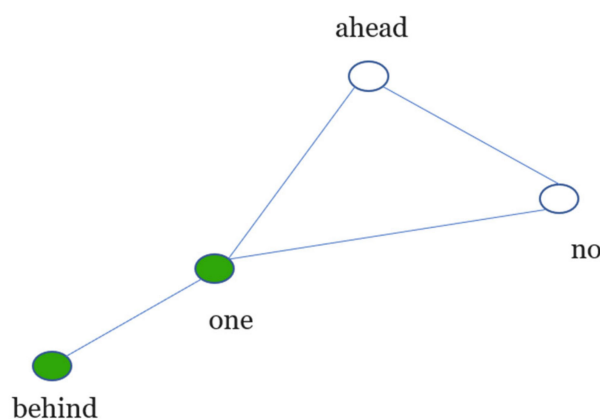


Figure 1. Word co-occurrence network of “No one behind, no one ahead”; the nodes in same colour belong to the same box.

For example, $N_b(l = 1) = 4$ and $N_b(l = \Delta + 1) = 1$ for the network of Figure 1. To obtain the $N_b(l = 2)$, we first compute a dual network (G') from the original (G) as follows: given a distance l ; two nodes i, j , in the dual network, are connected if the distance between l_{ij} is greater than or equal to l . For example, we start the procedure from the node “no”, see Figure 2; “no” and “behind” have a distance of two in

G , thus, they will be connected in G' . Next, the node “ahead” as the starting node is chosen—notice that the distance from it to “behind” is two—thus, a connection in G' will be drawn (see Figure 2).

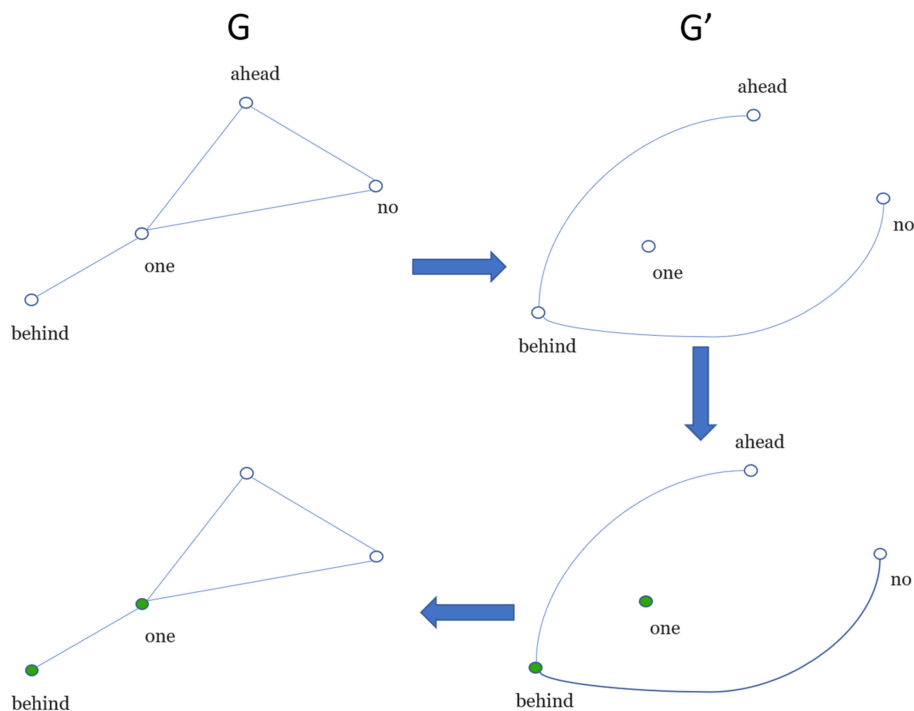


Figure 2. Covering of the network for a given box size ($l = 2$). The number of boxes in this network is $N_b(2) = 2$.

Then, the nodes of G' must be coloured following a single rule: two nodes directly connected will be painted different colours. The nodes of the resulting coloured dual network G' are mapped to original network G . The number of colours of G' represents the minimum number of boxes $N_b(l)$ of a given value of l to cover the network. The nodes of G , in the same colour, belong to the same box. The procedure described above is repeated until $l = \Delta + 1$. For more profound details of the box-covering algorithm, the reader is referred to the work of Chaoming, Lazaros [19]. Since l vs. $N_b(l)$ characterises the topology of the network, the area under the box-covering curve, l vs. $N_b(l)$ (AUB), was also included in the measures of the word co-occurrence network.

2.4. Robustness of Fractal Networks

Intentional network attacks are based on different centrality measures such as: the node degree or betweenness. They differ in the approach to compute those centrality measures such as: computing the global degree or betweenness, then performing the attack, or recomputing the centrality measure after a node is removed [20–23]. The fraction of nodes necessary to break down a fractal network (p_c) by a random attack are close to the total number of nodes; thus, these networks are extremely robust [24]. On the other hand, this robustness decreases drastically when the nodes with a high degree are selected to be removed [20,25]. This vulnerability to intentional attack relies on that a few nodes, with a high degree, maintain the connectivity of the network [26]. The robustness of each network is quantified by the size of the largest connected component C_c after removing a fraction p node from the network [20,24,26,27] when $C_c(p_c) \approx 0$ the network has been disintegrated. The value of p_c is low for fragile networks, and the opposite for robust networks.

Although the p_c value is useful for measuring the overall damage caused by the attack strategy, it does not reflect the damage of an individual node removal; for example, Figure 3 shows the plot of C_c vs. p , where the value of p_c is 0.5 and 0.49 for networks one and two, respectively.

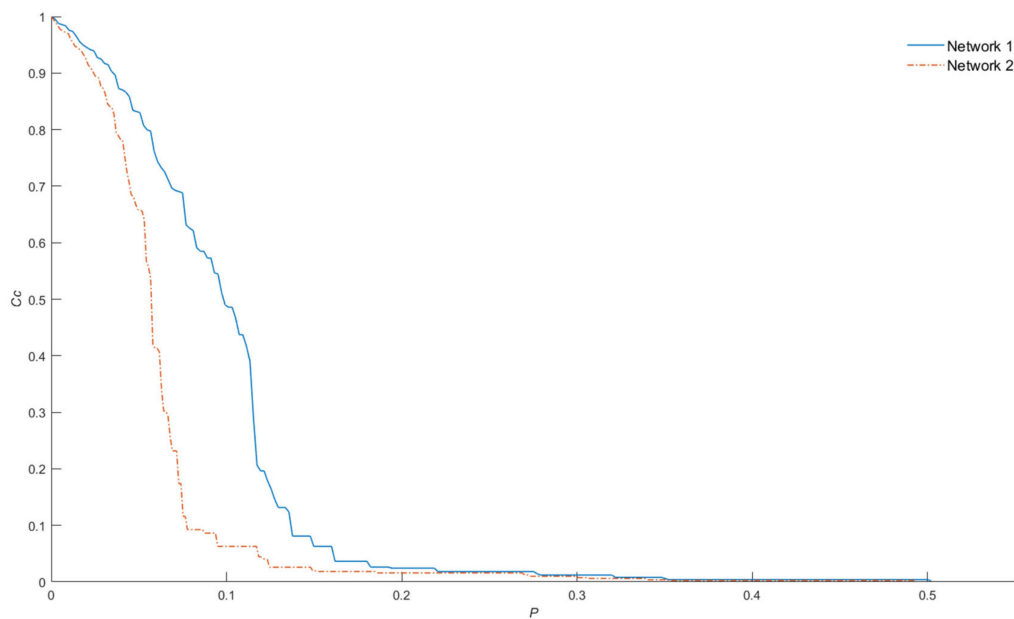


Figure 3. The damage of an individual node removal of network one and two. Although the p_c of networks one and two are 0.5 and 0.49, respectively, the area under the robustness curve reflects more precisely the vulnerability of the networks (0.0956 for network one and 0.060 for network 2).

This means that for both networks, it is necessary to remove approximately 50% of the nodes to disintegrate them in components that contain at most one node. Moreover, based on Figure 3, the removal of the nodes from network two causes more damage than the removal of those from network one. This damage can be quantified by computing the Area Under the Robustness Curve (AURC)—0.0956 for network one and 0.060 for network 2—to a higher the value, the higher the robustness of the network. The AURC of the attack performed by node degree was included as a measure of network robustness instead of p_c .

3. Materials and Methods

From seven Mexican writers —Juan José Arreola Zúñiga, Carlos Fuentes Macías, Jorge Ibarguengoitia Antillón, Carlos Monsiváis Aceves, José Emilio Pacheco Berny, Octavio Irineo Paz Lozano, and Alfonso Reyes Ocha—21 essays, 21 narratives (15 tales and six novels), and 21 research articles were the corpus for this research (see Table 1). Noticeably, some authors wrote titles classified as essays, tales, or novels, such as Carlos Fuentes, Jorge Ibarguengoitia, and José Emilio Pacheco. The essays, narratives, and research articles were published between 1911 and 2019. All the titles were obtained in an electronic format such as pdf and then converted to plain text.

The node degree (k), betweenness (b), shortest path length (spl), clustering coefficient (cc), and nearest neighbourhoods' degree (nnd), as well as more complex measures such as: the fractal dimension (d_b) obtained by the Equation (6), the complexity of a network $c(G)$ given by the Equation (8), the Area Under Box-covering (AUB), and the Area Under the Robustness Curve (AURC), were computed for each network of each title. Statistical analysis was carried out to select those measures that have a significant difference by literary genres and produce a better classification.

Then the Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), and Neural Network (NN) implemented in Weka [28] and fourth data mining views—described later and based on the measures mentioned above—were employed to classify the literary works. The hyperparameter optimisation of the data mining techniques was conducted by sequential model-based algorithm configuration [29,30]. The hidden layers and the nodes learning function of NN were 28 and sigmoid, respectively. The polynomial kernel was used in SVM, and all measures of the networks were

normalised before training and validating SVM and NN. The NN technique was used with a normal distribution to estimate the probabilities of the network measures. DT uses the C4.5 algorithm [31].

Table 1. The genre, number of titles, and primary author of the corpus.

Genre	Number of Titles	Primary Author
Essay	6	Alfonso Reyes Ochoa
Essay	3	Carlos Fuentes Macías
Essay	6	Carlos Monsiváis Aceves
Essay	6	Octavio Irineo Paz Lozano
Narrative (Tale)	2	Carlos Fuentes Macías
Narrative (Tale)	5	José Emilio Pacheco Berny
Narrative (Tale)	3	Jorge Ibarguengoitia Antillón
Narrative (Tale)	5	Juan José Arreola Zúñiga
Narrative (Novel)	1	Carlos Fuentes Macías
Narrative (Novel)	1	José Emilio Pacheco Berny
Narrative (Novel)	3	Jorge Ibarguengoitia Antillón
Narrative (Novel)	1	Juan José Arreola Zúñiga
Research Article	16	Several authors

The efficacy of each data mining technique and data mining views was validated by 5-fold cross-validation, comparing the Area under the Receiver Operating characteristic Curve (AROC). The AROC is useful to measure the performance of a data mining technique when the dataset is unbalanced [32]. Values of AROC closer to 1 mean a better classification than those closer to 0.5. This analysis shows the impact of data mining techniques and the measures on the classification of literary works. These results answer research question one (see Figure 4). Also, the accuracy of classification is presented as additional information that is computed as $(TP + True Negative (TN)) / (TP + False Positive (FP) + False Negative (FN) + True Negative (TN))$. The computation of AROC and accuracy are well-known for a two-class problem. Furthermore, for a multi-class problem, for each time one class could be considered as positive, then all the others as negative. This means that TP, TN, FP, and FN are calculated for each class. Therefore, a confusion matrix and AROC curve is obtained for each class (see [33,34] for more details).

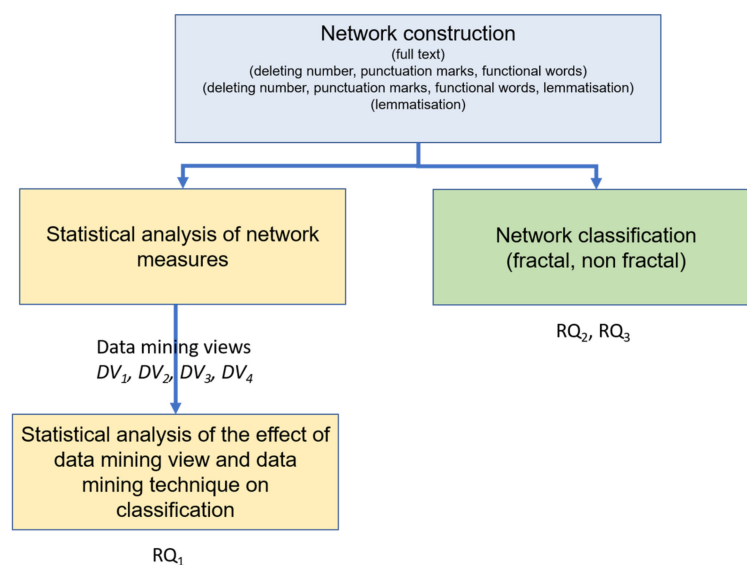


Figure 4. The experimental design followed to answer the research questions.

A set of word co-occurrence networks of each title was obtained and the first network was built using the full text. The second was obtained by deleting numbers and functional words.

A lemmatisation stage created the third after numbers and functional words deletion, and the fourth network was attained only through a lemmatisation stage (see Figure 4).

The networks were obtained by using the full text, by deleting numbers and functional words, by adding a lemmatisation stage after the numbers and functional words deletion, and through only a lemmatisation stage, are classified as fractal or non-fractal. Thus, research question two and three will be answered.

4. Results and Discussion

Tables 2–5 show the descriptive statistics by literary genre of the three types of networks—the first was built using the full text; the second was built by deleting numbers, punctuation marks, and functional words; the third was built by adding a lemmatisation stage; and the fourth was built through only a lemmatisation stage, denoted by subscripts *f*, *nf*, *l* and *ol*, respectively.

Table 2. Mean and standard deviation by genre of node degree (*k*), betweenness (*b*), shortest path length (*spl*), clustering coefficient (*cc*), nearest neighbourhoods’ degree (*nnd*), fractal dimension (*d_b*), complexity *c(G)*, the Area Under Box-covering (*AUB*), and the Area Under the Robustness Curve (*AURC*) of the networks built using the full text.

Genre	$K_f(\mu-\sigma)$	$b_f(\mu-\sigma)$	$spl_f(\mu-\sigma)$	$cc_f(\mu-\sigma)$	$nnd_f(\mu-\sigma)$	$d_{bf}(\mu-\sigma)$	$c(G)_f(\mu-\sigma)$	$AUB_f(\mu-\sigma)$	$AURC_f(\mu-\sigma)$
Essay	5.39–0.94	8249.16–4869.58	2.9–0.81	0.427–0.107	282.51–234.12	6.07–0.772	7.84×10^{-4} – 3.83×10^{-4}	1510.61–969.37	0.0154–0.0031
Narrative (Tale)	4.76–0.536	3868.33–1709.2	3.02–0.94	0.311–0.082	99.57–55.59	5.06–0.801	7.22×10^{-4} – 3.62×10^{-4}	685.86–315.90	0.0231–0.0067
Narrative (Novel)	6.81–1.01	13667.19–4796.52	2.78–0.061	0.577–0.088	559.34–233.84	7.42–0.644	6.78×10^{-4} – 2.48×10^{-4}	2631.5–1013.36	0.01456–0.0017
Research Article	5.80–0.94	5995.1–2013.42	3.00–0.092	0.374–0.065	153.31–73.31	5.43–0.539	3.72×10^{-4} – 2.39×10^{-4}	1068.4–379.76	0.0284–0.0071

Table 3. Mean and standard deviation by genre of node degree (*k*), betweenness (*b*), shortest path length (*spl*), clustering coefficient (*cc*), nearest neighbourhoods’ degree (*nnd*), fractal dimension (*d_b*), complexity *c(G)*, the Area Under Box-covering (*AUB*), and the Area Under the Robustness Curve (*AURC*) of the networks built by deleting numbers and functional words.

Genre	$k_{nf}(\mu-\sigma)$	$b_{nf}(\mu-\sigma)$	$spl_{nf}(\mu-\sigma)$	$cc_{nf}(\mu-\sigma)$	$nnd_{nf}(\mu-\sigma)$	$d_{bnf}(\mu-\sigma)$	$c(G)_{nf}(\mu-\sigma)$	$AUB_{nf}(\mu-\sigma)$	$AURC_{nf}(\mu-\sigma)$
Essay	3.763–1.067	15092.572–8418.110	5.370–0.911	0.052–0.037	12.101–9.638	2.057–0.352	1.5×10^{-5} – 1.32×10^{-5}	2299.330–1416.600	0.077–0.010
Narrative (Tale)	3.998–0.861	13572.998–7935.093	5.109–0.922	0.710–0.042	12.319–5.054	2.141–0.302	2.1×10^{-5} – 1.61×10^{-5}	2022.467–1173.400	0.083–0.018
Narrative (Novel)	4.703–0.425	16327.442–11950.302	4.460–0.249	0.097–0.029	17.797–5.444	2.390–0.160	1.8×10^{-5} – 0.77×10^{-5}	2656.333–1971.806	0.087–0.006
Research Article	3.483–1.054	15538.651–7392.827	5.770–1.03	0.037–0.024	10.648–9.539	1.940–0.383	1.2×10^{-5} – 0.59×10^{-5}	2339.476–1314.540	0.070–0.012

Table 4. Mean and standard deviation by genre of node degree (k), betweenness (b), shortest path length (spl), clustering coefficient (cc), nearest neighbourhoods' degree (nmd), fractal dimension (d_b), complexity $c(G)$, the Area Under Box-covering (AUB), and the Area Under the Robustness Curve ($AURC$) of the networks built by deleting numbers, functional words, and lemmatisation stage.

Genre	$k_l(\mu-\sigma)$	$b_l(\mu-\sigma)$	$spl_l(\mu-\sigma)$	$cc_l(\mu-\sigma)$	$nmd_l(\mu-\sigma)$	$d_{bl}(\mu-\sigma)$	$c(G)_l(\mu-\sigma)$	$AUB_l(\mu-\sigma)$	$AURC_l(\mu-\sigma)$
Essay	4.081–1.649	21871.411–7787.529	5.418–0.960	0.009–0.005	9.074–7.389	1.990–0.358	1.37×10^{-5} – 1.87×10^{-5}	2202.024–752.531	0.115–0.023
Narrative (Tale)	3.148–0.544	9538.206–4171.808	5.823–0.816	0.008–0.004	5.652–2.372	1.758–0.193	2.1×10^{-5} – 1.61×10^{-5}	1121.733–411.3845	0.003–0.0149
Narrative (Novel)	6.197–1.796	25284.524–7739.246	4.181–0.499	0.020–0.010	20.689–10.965	2.489–0.379	1.9×10^{-5} – 1.49×10^{-5}	2760.667–677.608	0.138–0.017
Research Article	4.849–0.691	11758.340–6325.385	4.296–0.352	0.207–0.010	11.169–2.802	2.234–0.155	2.8×10^{-5} – 1.85×10^{-5}	1272.857–540.06	0.138–0.016

Table 5. Mean, and standard deviation by genre of node degree (k), betweenness (b), shortest path length (spl), clustering coefficient (cc), nearest neighbourhoods' degree (nmd), fractal dimension (d_b), complexity $c(G)$, the Area Under Box-covering (AUB), and the Area Under the Robustness Curve ($AURC$) of the networks built only by lemmatisation stage.

Genre	$k_{ol}(\mu-\sigma)$	$b_{ol}(\mu-\sigma)$	$spl_{ol}(\mu-\sigma)$	$cc_{ol}(\mu-\sigma)$	$nmd_{ol}(\mu-\sigma)$	$d_{bol}(\mu-\sigma)$	$c(G)_{ol}(\mu-\sigma)$	$AUB_{ol}(\mu-\sigma)$	$AURC_{ol}(\mu-\sigma)$
Essay	5.377–0.94	2943.853–1852.075	2.915–0.081	0.413–0.093	280.854–232.41	6.041–0.775	7.85×10^{-4} – 3.84×10^{-4}	1520.381–972.673	0.0151–0.003
Narrative (Tale)	4.756–0.534	1324.018–628.419	3.033–0.095	0.289–0.06	99.135–55.332	5.019–0.798	7.24×10^{-4} – 3.63×10^{-4}	693.8–316.46	0.0242–0.007
Narrative (Novel)	6.794–1.014	5035.632–1864.014	2.793–0.062	0.507–0.093	555.98–231.929	7.383–0.673	6.79×10^{-4} – 2.49×10^{-4}	2644.833–1012.555	0.0151–0.002
Research Article	5.777–0.466	2096.629–725.505	3.015–0.095	0.349–0.05	152.065–72.851	5.378–0.559	3.59×10^{-4} – 2.42×10^{-4}	1078.357–379.765	0.0283–0.007

An Analysis of Variance (ANOVA) or a Kruskal–Wallis test—an ANOVA test carried out if the normality and homoscedasticity assumptions were valid for the given measure—was performed to select the measures of complex network that are influenced by essay, tale, novel, and research article genres. The one-way ANOVA conducted on the individual influence of essay, tale, novel, and research article on k_c , spl_f , cc_f , d_{bf} , $c(G)_f$, and $AURC_f$ shows significant effects: $F(3,59) = 12.81, p < 0.0001$; $F(3,59) = 15.039, p < 0.0001$; $F(3,59) = 14.77, p < 0.0001$; $F(3,59) = 19.27, p < 0.0001$; $F(3,59) = 6.40, p < 0.001$; and $F(3,59) = 22.35, p < 0.0001$. Similarly, a Kruskal–Wallis test shows a significant difference of the literary genres on nmd_f , AUB_f , b_f ; $H(3) = 29.44, p < 0.0001$; $H(3) = 27.98, p < 0.0001$; and $H(3) = 28.68, p < 0.0001$.

The one-way ANOVA conducted on the individual influence of essay, tale, novel, and research article on spl_{nf} , cc_{nf} , d_{bnf} , and $AURC_{nf}$ shows significant effects: $F(3,59) = 3.70, p = 0.016$; $F(3,59) = 6.17, p = 0.001$; $F(3,59) = 3.00, p \leq 0.037$; and $F(3,59) = 4.28, p = 0.008$. On the other hand, no effect on k_{nf} $F(3,59) = 2.65, p = 0.057$; nmd_{nf} $F(3,59) = 1.12, p = 0.347$; b_{nf} $F(3,59) = 0.227, p = 0.877$; $c(G)_{nf}$ $H(3) = 3.99, p = 0.262$; and AUB_{nf} $H(3) = 1.29, p = 0.731$ by genres were found. Although spl_{nf} , cc_{nf} , d_{bnf} , and $AURC_{nf}$ have a significant difference, they do not provide additional information—of those provided by the measures of full-text networks—to differentiate the genre. For example, spl_{nf} is only statistically different for the novel and tale (see Table 6). However, spl_f is statistically different for the novel, essay, and both the research article and tale. Thus, spl_{nf} , cc_{nf} , and d_{bnf} were not included in the set of measures to build data mining models. Table 6 summarises the significant statistical difference for spl_f and spl_{nf} .

Finally, the one-way ANOVA conducted on the individual influence of essay, tale, novel, and research article on spl_l and $AURC_l$ shows significant effects: $F(3,59) = 17.62, p < 0.0001$; $F(3,59) = 4.28, p = 0.008$. Similarly, a Kruskal–Wallis test shows a significant difference of k_l , cc_l , d_{bl} , $c(G)_l$, nmd_l , AUB_l , and b_l by genre: $H(3) = 32.9844, p < 0.0001$; $H(3) = 23.38, p < 0.0001$; $H(3) = 30.20, p < 0.0001$; $H(3) = 22.03, p < 0.0001$; $H(3) = 29.38, p < 0.0001$; $H(3) = 32.40, p < 0.0001, p < 0.0001$; and $H(3) = 32.64, p < 0.0001$.

Table 6. The subsets built using the significant statistical differences between slp_f and slp_{nf} induced by the novel, essay, research article, and tale. The value in the intersection of each row and column is the means of each measure for a given genre.

Genre	Subset 1	Subset 2	Subset 3
Novel- slp_f	2.78		
Essay- slp_f		2.90	
Research Article- slp_f			3.00
Tale- slp_f			3.02
Novel- slp_{nf}	4.46		
Essay- slp_{nf}	5.10	5.10	
Research Article- slp_{nf}	5.37	5.37	
Tale- slp_{nf}		5.77	

After these analyses, the spl_f , k_f , nnd_f , cc_f , b_f , db_f , $AURC_f$, AUB_f , $c(G)_f$, spl_l , k_l , nnd_l , cc_l , b_l , db_l , $AURC_l$, AUB_l , and $c(G)_l$ were selected to classify the genre of each literary work. This set of measures is a data mining view named DV_1 , and DV_1 was compared with a data mining view named DV_2 that contains all the measures computed on the three types of co-occurrence networks described previously. Also, a third data mining view named DV_3 , which contains only the measures spl , k , nnd , cc , and b obtained from the three types of co-occurrence networks, was tested to show that measures such as d_b , $c(G)$, AUB , and $AURC$ contribute to capturing the features of the literary genre. Since the influences of the data mining technique and data mining view on the AROC need to be tested, a two-way ANOVA is appropriate for this purpose, providing the data is normal and homoscedastic [32,35]. However, the AROC generated by our experiments does not meet these assumptions; thus, a Scheirer–Ray–Hare test [36,37] was used instead. A Scheirer–Ray–Hare test shows there is a significant difference among the AROC of the data mining views: $H(2) = 21.496$, $p < 0.001$, the data mining techniques: $H(3) = 84.79$, $p < 0.001$, and the interaction between both: $H(6) = 30.167$, $p < 0.001$. Figure 5 summarises the effect of both data mining view and data mining technique on AROC that are detailed below.

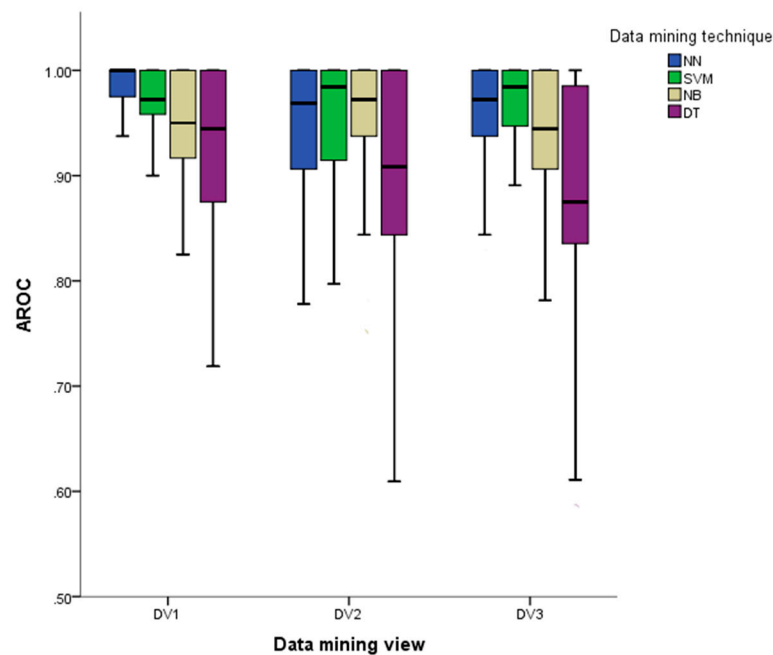


Figure 5. The effect of data mining view and data mining technique on Area under the Receiver Operating characteristic Curve (AROC).

A Kruskal–Wallis test shows that DV_1 , DV_2 , and DV_3 affect the median of the AROC:

$H(2) = 21.496, p < 0.001$. A posthoc Mann–Whitney test using a Dunn–Sidak adjustment [38] ($\alpha = 0.0169$) shows that the median of DV_1 ($Mdn = 0.975$) is higher than DV_2 ($Mdn = 0.968$)— $U(N_{DV1} = 400, N_{DV2} = 400) = 68704, z = -3.59, p < 0.001$ and DV_3 ($Mdn = 0.955$)— $U(N_{DV1} = 400, N_{DV2} = 400) = 66131, z = -4.388, p < 0.001$. Thus, the statistical analysis carried out on the measures of three types of networks is useful to select relevant measures that increase the AROC. No statistical difference was found between DV_2 and DV_3 , $U(N_{DV2} = 400, N_{DV2} = 400) = 78117, z = -0.59, p = 0.236$. This evidence suggests that well-known measures such as: node degree, shortest path length, betweenness, clustering coefficient, and the average of nearest neighbourhoods’ degree—used to build DV_3 —applied in the previous research to identify authors’ writing styles [3–8] are not enough to produce a higher AROC. On the other hand, more complex measures such as: $d_b, c(G), AUB$, and $AURC$ improve the classification.

Similarly, the Kruskal–Wallis test shows that the medians of the AROC obtained from NN, SVM, NB, and DT affect the AROC, $H(3) = 84.793$. A posthoc Mann–Whitney test using a Dunn–Sidak adjustment [38] ($\alpha = 0.0085$) shows that the median of both NN ($Mdn = 1.00$) and SVM ($Mdn = 0.975$) were higher than those of NB ($Mdn = 0.968$)—see the corresponding row and column of Table 7 for the result of the pair-wise test e.g., row NB and column NN show a significant difference: $U(N_{NN} = 300, N_{NB} = 300) = 36784.5, z = -3.995, p < 0.0001$ — and DT ($Mdn = 0.911$). No statistical difference between NN ($Mdn = 1.00$) and SVM ($Mdn = 0.975$) was found.

Table 7. Pair-wise Mann–Whitney test using a Dunn–Sidak adjustment ($\alpha = 0.0085$) among data mining techniques. The intersection of a row and a column presents the result of the test between the two data mining techniques.

	NN	SVM	NB	DT
NN	—			
SVM	$U(N_{NN} = 300, N_{SVM} = 300) = 41859, z = -1.55, p < 0.120$	—		
NB	$U(N_{NN} = 300, N_{NB} = 300) = 36784.5, z = -3.995, p < 0.0001$	$U(N_{SVM} = 300, N_{NB} = 300) = 35311, z = -4.749, p < 0.0001$	—	
DT	$U(N_{NN} = 300, N_{DT} = 300) = 29119, z = -7.816, p < 0.0001$	$U(N_{NN} = 300, N_{DT} = 300) = 30523, z = -6.989, p < 0.0001$	$U(N_{NN} = 300, N_{DT} = 300) = 35438.5, z = -4.588, p < 0.0001,$	—

Then a significant difference between NB ($Mdn = 0.968$) and DT ($Mdn = 0.911$), was found. These results suggest that the DV_1 and the use of NN or SVM produce statistically equal values of AROC. The accuracy of NN and SVM based on DV_1 are 0.93 and 0.90, respectively, based on DV_1 .

To support the conjecture that deleting number, punctuation, and functional words do not have a significant effect on the AROC, the models of NN based on DV_1 and the fourth data mining view named DV_4 , which contain the measures from the networks built using the full text ($spl_f, k_f, nnd_f, cc_f, b_f, db_f, AURC_f, AUB_f$, and $c(G)_f$) and those from networks built using only a lemmatisation stage ($spl_{ol}, k_{ol}, nnd_{ol}, cc_{ol}, b_{ol}, db_{ol}, AURC_{ol}, AUB_{ol}$, and $c(G)_{ol}$), were compared. The Mann–Whitney test shows no statistical difference— $U(N_{DV1} = 100, N_{DV2} = 100) = 4793, z = -0.631, p = 0.528$ —between the AROC of DV_1 ($Mdn = 1$) and DV_4 ($Mdn = 0.98$). The accuracy of DV_1 and DV_4 is 0.93 for both. Thus, the deletion of the number and punctuation marks is not useful to reveal stylistic attributes by genre as lemmatisation does. Furthermore, all these stages together modify the network fractality, as the evidence presented later suggests. The accuracy of the NN model based on DV_1, DV_2, DV_3 , and DV_4 are 0.93, 0.90, 0.89, and 0.93, respectively.

To classify each network as fractal or non-fractal, the Akaike Information Criterion (AIC) [39] were computed for the networks based on the full text. The second network was obtained by deleting numbers, punctuation marks, and functional words. The third was created by adding a lemmatisation stage, and the fourth was attained only through a lemmatisation. The AIC is useful to classify networks

as fractal and non-fractal [40]. To select the better mathematical model, first the AIC for power (denoted by subscript P) and exponential (denoted by subscript E) models—Equations (5) and (7)—were computed, then the minimum value is chosen (AIC_{\min}). ΔAIC_i was computed by $AIC_i - AIC_{\min}$, where i is the AIC of power or exponential models. The AIC's rule of thumb is that the two models are statistically different if ΔAIC is greater than two, thus, the model with $\Delta AIC = 0$ should be selected [41,42]. Table S2 of the supplementary material shows that the difference between ΔAIC_P and ΔAIC_E for about 87% of the full word co-occurrence network is higher than two; thus, the mathematical model for the relation l vs. $N_b(l)$ computed by the box-covering algorithm of these networks is the power model (see Equation (5)). Although for 13% of the networks, a model cannot be selected feasibly based on ΔAIC , the power model obtained the least value. Thus, most of the full word co-occurrence networks of literary works are fractal. This result supports the fractality founded in other languages and English literature by different mathematical analyses [43–46]. Noticeably, selecting the better model based on the adjusted coefficient of determination (R^2) is rather difficult.

Similarly, Table S3 of the supplementary material shows that the difference between ΔAIC_P and ΔAIC_E for about 89% of the word co-occurrence networks—built by deleting numbers, punctuation marks, and functional words—suggests they are fractal; 2% were classified as exponential, and 9% were undetermined (since $\Delta AIC \leq 2$). However, adding a lemmatisation stage to the previous ones dilutes the fractality (25.3% are fractal, 33.3% are exponential, and 41.3 are undetermined), see Table S4. The lemmatisation stage alone preserves the fractality of the full-text networks (87% are fractals, and 13% are undetermined); see Tables S2 and S5, which show no difference between the AROC curve of the classification of literary works according to their genre. Note that the lemmatisation stage preserves the original fractality of the networks. Thus, this supports the conjecture that lemmatisation is a kind of renormalisation of a complex network that preserves the fractality. This paves the way to compare this linguistic renormalisation with that introduced by Song, Havlin [16].

5. Conclusions

This research aims at showing that measures of the word co-occurrence network of literary works—by Mexican writers—classifies them according to the literary genre. The local measures—such as: node degree, the average of nearest neighbourhoods' degree, and global measures using shortest path length, betweenness, clustering coefficient, and the average of nearest neighbourhoods' degree—widely used in the previous research to identify authors' writing styles, produces acceptable values of AROC classification. However, more elaborate measures using fractal dimension, complexity, the AUB, and the AURC show an improvement of AROC. These measures capture the topology based on the minimum number of boxes to cover the network, the robustness, and the complexity measured by structural entropy and density. Precisely 87% of the full word co-occurrence networks were classified as a fractal. Thus, those findings support the conjecture that fractality occurs in the literary works of Mexican writers, as was previously reported by their English-speaking counterparts. Also, the empirical evidence suggests that the lemmatisation of literary works is a renormalisation stage that preserves the original text fractality. On the contrary, the deletion of numbers, punctuation marks, and functional works, as well as lemmatisation, dilute the fractality. The number of literary works included in this study limit the generalisation of this conjecture. Also, it would be interesting for future research directions to compare the renormalisation induced by a lemmatisation stage—linguistic renormalisation—to renormalisation of networks based on the box-covering algorithm.

Supplementary Materials: The following are available online at <http://www.mdpi.com/1099-4300/22/8/904/s1>, Table S1 Title, primary author and the gender of the literary works, Table S2, The adjusted determination coefficient, AIC, ΔAIC and classification for the networks based the full text, Table S3 The adjusted determination coefficient R^2 , AIC, ΔAIC and classification for the networks obtained by deleting numbers, punctuation marks and functional words, Table S4 The adjusted determination coefficient R^2 , AIC, ΔAIC and classification for the networks obtained by deleting numbers, punctuation marks, functional words and a lemmatisation stage, and Table S5 The adjusted determination coefficient, AIC, ΔAIC and classification for the networks obtained only by a lemmatisation stage.

Funding: This research was funded by Instituto Politécnico Nacional, grant number SIP20200169.

Acknowledgments: An especial mention is given to Alejandro Rosas who provide the corpus of Mexican writers used for experiments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fang, Y.; Wang, Y. Quantitative Linguistic Research of Contemporary Chinese. *J. Quant. Linguist.* **2018**, *25*, 107–121. [[CrossRef](#)]
2. Cong, J.; Liu, H. Approaching human language with complex networks. *Phys. Life Rev.* **2014**, *11*, 598–618. [[CrossRef](#)] [[PubMed](#)]
3. Amancio, D.R.; Oliveira, O.N., Jr.; Costa, L.d.F. Structure–semantics interplay in complex networks and its effects on the predictability of similarity in texts. *Phys. A Stat. Mech. Appl.* **2012**, *391*, 4406–4419. [[CrossRef](#)]
4. Akimushkin, C.; Amancio, D.R.; Oliveira, O.N. On the role of words in the network structure of texts: Application to authorship attribution. *Phys. A Stat. Mech. Appl.* **2018**, *495*, 49–58. [[CrossRef](#)]
5. Mehri, A.; Darooneh, A.H.; Shariati, A. The complex networks approach for authorship attribution of books. *Phys. A Stat. Mech. Appl.* **2012**, *391*, 2429–2437. [[CrossRef](#)]
6. Darooneh, A.H.; Shariati, A. Metrics for evaluation of the author’s writing styles: Who is the best? *Chaos Interdiscip. J. Nonlinear Sci.* **2014**, *24*, 033132.
7. Machicao, J.; Corrêa Jr, E.A.; Miranda, G.H.; Amancio, D.R.; Bruno, O.M. Authorship attribution based on Life-Like Network Automata. *PLoS ONE* **2018**, *13*, e0193703. [[CrossRef](#)]
8. Stanisiz, T.; Kwapien, J.; Drożdż, S. Linguistic data mining with complex networks: A stylometric-oriented approach. *Inf. Sci.* **2019**, *482*, 301–320. [[CrossRef](#)]
9. Amancio, D.R.; Nunes, M.D.; Oliveira Jr, O.N.; Pardo, T.A.; Antiqueira, L.; Costa, L.D. Using metrics from complex networks to evaluate machine translation. *Phys. A Stat. Mech. Appl.* **2011**, *390*, 131–142. [[CrossRef](#)]
10. Ferraz de Arruda, H.; Nascimento Silva, F.; Queiroz Marinho, V.; Raphael Amancio, D.; da Fontoura Costa, L. Representation of texts as complex networks: A mesoscopic approach. *J. Complex Netw.* **2017**, *6*, 125–144. [[CrossRef](#)]
11. De Arruda, H.F.; Marinho, V.Q.; Costa, L.D.; Amancio, D.R. Paragraph-based representation of texts: A complex networks approach. *Inf. Process. Manag.* **2019**, *56*, 479–494. [[CrossRef](#)]
12. Kim, J.; Wilhelm, T. What is a complex graph? *Phys. A Stat. Mech. Appl.* **2008**, *387*, 2637–2652. [[CrossRef](#)]
13. Van Steen, M. *Graph Theory and Complex Networks: An Introduction*; Cambridge University Press: Cambridge, UK, 2010.
14. Estrada, E. *The Structure of Complex Networks: Theory and Applications*; Oxford University Press: Oxford, UK, 2012.
15. Gallos, L.K.; Song, C.; Makse, H.A. A review of fractality and self-similarity in complex networks. *Phys. A Stat. Mech. Appl.* **2007**, *386*, 686–691. [[CrossRef](#)]
16. Song, C.; Havlin, S.; Makse, H.A. Origins of fractality in the growth of complex networks. *Nat. Phys.* **2006**, *2*, 275–281. [[CrossRef](#)]
17. Lei, M.; Liu, L.; Wei, D. An Improved Method for Measuring the Complexity in Complex Networks Based on Structure Entropy. *IEEE Access* **2019**, *7*, 159190–159198. [[CrossRef](#)]
18. Scott, J. Social network analysis. *Sociology* **1988**, *22*, 109–127. [[CrossRef](#)]
19. Song, C.; Gallos, L.K.; Havlin, S.; Makse, H.A. How to calculate the fractal dimension of a complex network: The box covering algorithm. *J. Stat. Mech. Theory Exp.* **2007**, *2007*, P03006. [[CrossRef](#)]
20. Holme, P.; Kim, B.J.; Yoon, C.N.; Han, S.K. Attack vulnerability of complex networks. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **2002**, *65 Pt 2*, 056109. [[CrossRef](#)]
21. Callaway, D.S.; Newman, M.E.; Strogatz, S.H.; Watts, D.J. Network Robustness and Fragility: Percolation on Random Graphs. *Phys. Rev. Lett.* **2000**, *85*, 5468–5471. [[CrossRef](#)]
22. Cohen, R.; Erez, K.; Ben-Avraham, D.; Havlin, S. Resilience of the Internet to Random Breakdowns. *Phys. Rev. Lett.* **2000**, *85*, 4626–4628. [[CrossRef](#)]
23. Cohen, R.; Erez, K.; Ben-Avraham, D.; Havlin, S. Breakdown of the Internet under Intentional Attack. *Phys. Rev. Lett.* **2001**, *86*, 3682–3685. [[CrossRef](#)]
24. Albert, R.; Jeong, H.; Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **2000**, *406*, 378. [[CrossRef](#)] [[PubMed](#)]

25. Gallos, L.K.; Cohen, R.; Argyrakis, P.; Bunde, A.; Havlin, S. Stability and Topology of Scale-Free Networks under Attack and Defense Strategies. *Phys. Rev. Lett.* **2005**, *94*, 188701. [[CrossRef](#)] [[PubMed](#)]
26. Gallos, L.K.; Cohen, R.; Argyrakis, P.; Bunde, A.; Havlin, S. *Fractal and Transfractal Scale-Free Networks, in Encyclopedia of Complexity and Systems Science*; Meyers, R.A., Ed.; Springer: New York, NY, USA, 2009; pp. 3924–3943.
27. Iyer, S.; Killingback, T.; Sundaram, B.; Wang, Z. Attack Robustness and Centrality of Complex Networks. *PLoS ONE* **2013**, *8*, e59613. [[CrossRef](#)]
28. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [[CrossRef](#)]
29. Hutter, F.; Hoos, H.H.; Leyton-Brown, K. *Sequential Model-Based Optimisation for General Algorithm Configuration*; Springer: Berlin/Heidelberg, Germany, 2011.
30. Kotthoff, L.; Thornton, C.; Hoos, H.H.; Hutter, F.; Leyton-Brown, K. Auto-WEKA: Automatic Model Selection and Hyperparameter Optimization in WEKA. In *Automated Machine Learning: Methods, Systems, Challenges*; Hutter, F., Kotthoff, L., Vanschoren, J., Eds.; Springer International Publishing: Cham, Germany, 2019; pp. 81–95.
31. Quinlan, J.R. *C4. 5: Programs for Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2014.
32. Ramirez-Arellano, A.; Bory-Reyes, J.; Hernandez-Simon, L.M. Statistical Entropy Measures in C4.5 Trees. *Int. J. Data Warehous. Min. (IJDWM)* **2018**, *14*, 1–14. [[CrossRef](#)]
33. Tharwat, A. Classification assessment methods. *Appl. Comput. Inform.* **2018**. [[CrossRef](#)]
34. Hand, D.J.; Till, R.J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Mach. Learn.* **2001**, *45*, 171–186. [[CrossRef](#)]
35. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
36. Scheirer, C.J.; Ray, W.S.; Hare, N. The Analysis of Ranked Data Derived from Completely Randomised Factorial Designs. *Biometrics* **1976**, *32*, 429–434. [[CrossRef](#)]
37. Dytham, C. *Choosing and Using Statistics: A Biologist's Guide*; Wiley: Hoboken, NJ, USA, 2011.
38. Ennos, A.R. *Statistical and Data Handling Skills in Biology*; Pearson/Prentice Hall: Upper Saddle River, NJ, USA, 2007.
39. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [[CrossRef](#)]
40. Ramirez-Arellano, A. Students learning pathways in higher blended education: An analysis of complex networks perspective. *Comput. Educ.* **2019**, *141*, 103634. [[CrossRef](#)]
41. Burnham, K.P.; Anderson, D.R. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociol. Methods Res.* **2004**, *33*, 261–304. [[CrossRef](#)]
42. Burnham, P.K.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*; Springer: Berlin/Heidelberg, Germany, 2002; Volume 67.
43. Andres, J. On a Conjecture about the Fractal Structure of Language. *J. Quant. Linguist.* **2010**, *17*, 101–122. [[CrossRef](#)]
44. Hřebíček, L.K. Fractals in language. *J. Quant. Linguist.* **1994**, *1*, 82–86.
45. Glattre, H.R.; Glattre, E. Finding Fractal Networks in Literature. *Nonlinear Dyn. Psychol Life Sci.* **2018**, *22*, 263–282.
46. Kohler, R. Are there fractal structures in language? Units of measurement and dimensions in linguistics. *J. Quant. Linguist.* **1997**, *4*, 122–125. [[CrossRef](#)]

