UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Predição de sucesso de livros por meio de uma abordagem de redes complexas

Alana Isabelle Uchiyama

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Alana Isabelle Uchiyama

Predição de sucesso de livros por meio de uma abordagem de redes complexas

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Diego Raphael Amancio

Versão original

São Carlos

2022

Alana Isabelle Uchiyama

Prediction of books success through complex networks approach

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Diego Raphael Amancio

Original version

São Carlos

2022

RESUMO

Uchiyama, A.I **Predição de sucesso de livros por meio de uma abordagem de redes complexas**. 2022. 48p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

O fenômeno de um livro entrar para a lista dos mais vendidos continua sendo um mistério, mesmo com a existência de diversas pesquisas sobre o tema. Este trabalho tem como objetivo extrair padrões a partir do contexto e do desdobramento da história que consigam predizer o que leva um livro ao sucesso. Para esse propósito, foi empregada uma abordagem de redes complexas mesoscópicas para a representação do texto, uma vez que essa técnica preserva a sequência dos acontecimentos e faz relações entre assuntos já mencionados no enredo. A avaliação da metodologia adotada ocorreu por meio da análise dos resultados de algoritmos supervisionados de classificação, além da comparação do desempenho dos modelos adotando outras técnicas de processamento de texto.

Palavras-chave: Inteligência artificial. Redes complexas. Processamento de texto. Algoritmo supervisionado.

ABSTRACT

Uchiyama, A.I **Prediction of books success through complex networks approach**. 2022. 48p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

The phenomenon of a book becoming a best seller remains a mystery, even with the existence of several researches on the subject. The present work aims to extract patterns from the context and story unfolding that can predict what makes a book successful. For this purpose, an approach of complex mesoscopic networks was used for text representation, since this technique preserves the sequence of events and makes relationships between the subjects already mentioned in the narrative. Methodology's evaluating ran through the analysis of supervised classification algorithms metrics, in addition to the comparison of models performance adopting other text processing techniques.

Keywords: Artificial intelligence. Complex networks. Text processing. Supervised classification algorithms.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Introdução e justificativa

A escrita desde muito tempo é uma forma de contar histórias e transmitir conhecimento, sendo os livros uma das maneiras mais populares de consumir essas narrativas e experiências. Para mostrar que esse hábito de consumo continua em alta, no ano de 2015 cerca de 2,7 bilhões de livros foram vendidos no mercado norte americano (WANG *et al.*, 2019). Esse mercado não é apenas expressivo em volume de vendas, mas financeiramente também, em 2019 a indústria de publicação de livros nos Estados Unidos obteve uma receita líquida de 25,93 bilhões de dólares americanos (STATISTA,).

No entanto, são poucos os livros e os autores que ingressam no seleto grupo de obras que obtiveram sucesso de vendas. Dos mais de 100 mil novos títulos impressos nos EUA em 2015, apenas por volta de 4 mil venderam mais de mil cópias em um ano e cerca de 500 livros entraram para as listas de mais vendidos do *New York Times* (WANG *et al.*, 2019).

Apesar de existirem diversas pesquisas estudando o que leva um livro ao sucesso, inúmeros fatores podem influenciar para esse resultado, como por exemplo o estilo da escrita do autor (ASHOK; FENG; CHOI, 2013), as críticas publicadas sobre o livro (CLEMENT; PROPPE; ROTT, 2007), o nível de exposição e rede de contatos do autor (NAKAMURA, 2013; CARMI; OESTREICHER-SINGER; SUNDARARAJAN, 2012), a estratégia de publicidade (SHEHU *et al.*, 2013) e entre outros.

Prever se o lançamento de um livro vai ser sucesso e consequentemente trazer resultados financeiros são os principais motivadores para as editoras tomarem a decisão de publicar ou não uma obra. Ainda assim, é um trabalho difícil, dado que atualmente as editoras definem a publicação de um livro com base no sucesso do autor em obras anteriores, o apelo do tema e o comportamento de venda de livros similares (WANG *et al.*, 2019).

Dado esse cenário e o sucesso de aplicação de redes complexas ao modelar textos, como por exemplo análise de sentimento (FELDMAN, 2013) e estilometria (AMANCIO, 2015b), isso torna a abordagem por redes complexas uma alternativa interessante para o problema apresentado. Estudos recentes de redes linguísticas, como o proposto por Arruda *et al.* (2018) apresentou que modelar textos utilizando uma escala textual maior, ou seja, ao invés de considerar palavras como unidade, utilizar fragmentos maiores (sentenças ou parágrafos, por exemplo) possibilita entender e visualizar o desdobramento das histórias.

Com o intuito de analisar o fenômeno de vendas de um livro por outras perspectivas, uma abordagem de redes complexas é proposta com a finalidade de estudar a influência

que a estrutura semântica e a evolução temporal dos acontecimentos têm sobre o sucesso de vendas dos livros, através da perspectiva mesoscópica (ARRUDA *et al.*, 2018). O conhecimento do impacto dos fatores estudados no sucesso de uma história poderá ser uma poderosa ferramenta para auxiliar a tomada de decisão das editoras.

1.2 Objetivos

Este trabalho tem como finalidade identificar padrões capazes de classificar um livro de sucesso pela abordagem de redes mesoscópicas. O conceito de sucesso será dado pela presença do título na lista de *bestsellers* da revista *Publishers Weekly*. Diante dos desafios de trabalhar com classificação de textos longos, foram propostos os seguintes objetivos:

1. Transformar livros em redes mesoscópicas e através das métricas extraídas dessas redes, desenvolver e avaliar modelos de classificação cuja a variável resposta é o sucesso de vendas.

2. Comparar e avaliar desempenho dos classificadores treinados a partir de variáveis extraídas de redes mesoscópicas e de técnicas de processamento de linguagem natural.

1.3 Organização do trabalho

Sobre a organização desse tranalho, o Capítulo 2 contém o referencial teórico introduzindo conceitos gerais sobre redes complexas. No Capítulo 3, foi apresentado a metodologia proposta com o processamento de texto realizado, detalhes da transformação do texto em redes mesoscópicas e as métricas de avaliação dos modelos. O Capítulo 4 abrange análise dos resultados para cada experimento realizado em conjunto com a avaliação dos algoritmos de classificação testados. O Capítulo 5 compreende a conclusão dos resultados alcançados e possibilidades futuras.

2 FUNDAMENTOS E TRABALHOS RELACIONADOS

2.1 Redes

2.1.1 Definição de redes

Uma rede é caracterizada por ser um conjunto de itens interligados, no qual os itens são denominados vértices ou nós, já as conexões existentes entre eles são chamadas de arestas (Figura 1). Os sistemas representados por uma rede também podem ser chamados de grafos e estão presentes de várias formas no mundo, como por exemplo a internet, redes sociais, redes neurais e redes de relações de negócios entre empresas.

Além disso, há várias maneiras dos vértices e arestas se relacionarem em uma rede, considerando propriedades tal como intensidade e direção. Se existe peso nas arestas, como em uma rede na qual retrata o quão bem as pessoas se conhecem, então se caracteriza um grafo ponderado. Quando as arestas apontam apenas para uma direção conectando um par de vértices, ou seja, o grafo é composto por vértices direcionados, é definido como grafo orientado ou digrafo. Um grafo onde as arestas são e-mails trocados entre indivíduos e as mensagens possuem apenas uma direção, é um exemplo de digrafo (NEWMAN, 2003).

De forma matemática, uma rede G é definida por $G = (V, E)$ composta por um conjunto $V = \{v_1, v_2, \ldots, v_M\}$ de vértices e por um conjunto $E = \{e_1, e_2, \ldots, e_E\}$ de arestas. A matriz A, simétrica, não ponderada e de tamanho M, indica a conectividade dos vértices, ou seja, o vértice v_i está interligado ao vértice v_j quando $a_{ij} = 1$.

A rede ponderada é representada por $G^w = (V, E, W)$, onde V é o conjunto de vértices, E é o conjunto de arestas e $W = \{w_1, w_2, \ldots\ w_E\}$ é o conjunto de valores reais

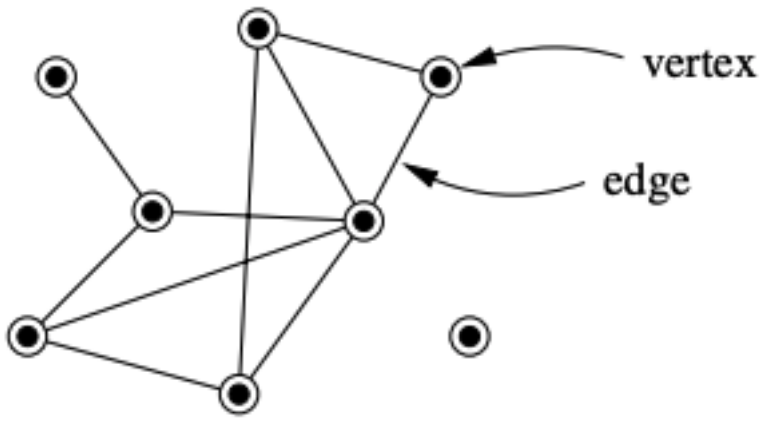

Figura 1 – Estrutura de uma rede contendo vértices e arestas (Extraído de M.E.J. Newman, 2003.

do peso das arestas.

No caso das redes orientadas, a direção está indicada nas matrizes A e W (se for uma rede ponderada), dessa forma o elemento a_{ij} indica se há a conexão $v_i \rightarrow \mathrm{v}_j$ e o elemento w_{ij} guarda o peso do arco $(v_i \rightarrow \mathrm{v}_j)$ (AMANCIO, 2013).

2.1.2 Redes complexas

O estudo de redes complexas se originou a partir da teoria dos grafos, tal teoria teve seu início com a solução de Leonhard Euler para o conhecido problema das pontes de Konigsberg (AMARAL; OTTINO, 2004). O modelo que guiou por décadas o campo de redes complexas foi o de grafo aleatório, inicialmente estudado pelos matemáticos Paul Erdos e Afred Renyi. Um grafo aleatório seguindo o modelo Erdos-Renyi possui M vértices, com arestas conectando um par de vértices a uma probabilidade p e distribuídas de forma aleatória.

Entretanto, com o avanço dos estudos pode-se perceber que redes reais apresentam características que não são bem descritas por modelos de conexões aleatórias. Assim sendo, outras abordagens e métricas foram propostas com o intuito de explicar a complexidade de sistemas reais (ALBERT; BARABASI, 2002). Existem três conceitos que ocupam relevância na pesquisa contemporânea de redes complexas, são eles: redes pequeno mundo, aglomeração e a distribuição de graus.

Redes pequeno mundo (ou *small worlds*) definem que independentemente do tamanho da rede, existe um caminho de baixo custo conectando quaisquer dois vértices, em outras palavras, a quantidade de vértices intermediários entre um par de vértices é pequena (AMARAL; OTTINO, 2004; ALBERT; BARABASI, 2002). O conceito de *small worlds* está bem representado no famoso experimento "seis graus de separação", no qual conclui que a distância de quaisquer duas pessoas são seis indivíduos intermediários (WATTS, 2003).

Em redes sociais, é comum notar a formação de grupos (ou *clusters*) tal como em um círculo de amigos em que todos se conhecem. O coeficiente de aglomeração (*clustering coefficient*) que será abordado com mais detalhes no item 2.3, é capaz de quantificar a tendência de aglomeração. Em redes complexas, o coeficiente de aglomeração na maior parte dos casos é significativamente maior em comparação a grafos aleatórios de mesmo número de vértices e arestas, assim indicado por Watts e Strogatz (ALBERT; BARABASI, 2002).

Por fim, o grau é denominado pela quantidade de arestas conectadas a um vértice (NEWMAN, 2003), logo a distribuição de graus é a distribuição do número de conexões de um vértice em uma rede. Grafos aleatórios possuem uma distribuição de graus sendo uma distribuição de Poisson. No entanto, ao analisar grandes redes como a internet, observa-se

uma distribuição de graus seguindo uma lei de potência $P(k) \sim k^{-\gamma k}$ e são chamadas de redes livre de escala, ou *scale-free* (ALBERT; BARABASI, 2002). O modelo de rede de escala livre proposto por Barabási e Albert também aponta que em um cenário de crescimento, os novos vértices preferencialmente se ligarão a vértices que já possuem uma grande quantidade de conexões, ou *hubs* (AMARAL; OTTINO, 2004).

Em resumo, dado os conceitos mencionados, pode-se dizer que as redes complexas são caracterizadas pela distância entre dois vértices quaisquer ser representada por uma quantidade razoavelmente pequena de vértices (redes pequeno mundo), pela tendência em formar grupos fortemente interligados (alto coeficiente de aglomeração) e a presença de *hubs* devido ao comportamento de redes livres de escala.

2.2 Representação de textos em redes

A língua humana é um sistema complexo, por esse motivo, redes complexas são capazes de representá-la de acordo com sua multiplicidade e assim através dos modelos obter métricas quantitativas. Não existe um modelo único que consiga exprimir a natureza multifacetada da língua humana, por isso há várias abordagens de redes complexas para cada subsistema da língua, de acordo com aspectos particulares, como a semântica ou sintaxe.

As redes linguísticas possibilitam modelar subsistemas de inventários de unidades linguísticas (palavras, morfemas e fonemas por exemplo), tal como dicionários. A relação entre as unidades linguísticas pode se dar pela semântica, usando recursos de associação de palavras como *WordNet*, dicionário de sinônimos. Dessa forma, os vértices são as unidades linguísticas que são conectados por regras semânticas como arestas.

Já os modelos orientados pelas relações de dependência sintática, são chamados de redes sintáticas. Nesse caso os vértices são palavras e as arestas se dão por meio da existente relação de dependência sintática entre os vértices em uma sentença.

Para representar subsistemas que constituem o uso da língua, os modelos mais adequados são os que se baseiam em dados de linguagem de ocorrência natural. A rede de co-ocorrência se destaca dentro dessa abordagem, capaz de expressar a linearidade das palavras em uma sentença por meio da conexão de palavras adjacentes (CONG; LIU, 2014). Posteriormente, a rede mesoscópica foi desenvolvida com a finalidade de modelar fragmentos maiores de texto e assim identificar linearidade e o desdobramento temporal dos textos. Por essa razão esses dois modelos de ocorrência natural serão mais detalhados a seguir.

2.2.1 Redes de co-ocorrência

A rede de co-ocorrência é estruturada a partir de textos, no qual os vértices são as palavras e as arestas são formadas conectando palavras que são imediatamente adjacentes no texto. Por ser um processo de construção simples e ao mesmo tempo conseguir extrair aspectos semânticos e sintáticos de um texto, sua aplicação obteve sucesso em diversos problemas de processamento de linguagem natural, tanto em análise de sentimento e estilometria mencionados anteriormente, como também em detecção de autoria (AMANCIO, 2015a) e classificação de texto (ARRUDA; COSTA; AMANCIO, 2016).

Além disso, a rede de co-ocorrência pode ser considerada uma aproximação de uma rede sintática, pois é capaz de capturar a maior parte das ligações sintáticas, uma vez que a maioria dessas ligações ocorrem entre palavras vizinhas (AMANCIO, 2015b).

Matematicamente, a rede de co-ocorrência é definida por um conjunto de vértices $V = \{v_1, v_2, \ldots, v_n\}$ e um conjunto E de arestas. O conjunto E é representado como uma matriz W ponderada e não simétrica, onde W é uma matriz quadrada de tamanho m, sendo m o número distinto de palavras no texto após o pré-processamento. Na matriz W, os elementos w_{ij} indicam quantas vezes a palavra v_j aparece logo após a palavra v_i, ou seja, w_{ij} mostra a intensidade em que o vértice v_i está conectado ao vértice v_j ($v_i \rightarrow$v$_j$). Nesse método também existe uma matriz de adjacência A em que seus elementos a_{ij} equivalem a 1 quando as palavras representadas pelos vértices v_i e v_j ocorrem adjacentes ao menos uma vez no texto, senão $a_{ij} = 0$ (AMANCIO *et al.*, 2011).

2.2.2 Abordagem mesoscópica para representação de textos

Apesar das redes de co-ocorrência serem amplamente utilizadas em problemas de processamento de linguagem natural, quando se trata de capturar características sobre o conteúdo (tópicos e subtópicos) e a evolução temporal dos textos, esse modelo tem limitações. Isso ocorre pois as redes de co-ocorrência tratam de uma abordagem microscópica que identificam majoritariamente atributos sintáticos e raramente apresentam estrutura de comunidade (ARRUDA *et al.*, 2018). Dado esse cenário, os estudos de redes linguísticas se direcionaram para definir uma unidade representacional básica com escala textual maior, ou seja, modelando textos nos quais os vértices são representados por sentenças, parágrafos ou fragmentos maiores, apresentando assim uma perspectiva mesoscópica.

Um modelo de redes mesoscópicas que captura características de conteúdo semântico em textos e explicita o desdobramento ao longo do tempo foi proposto em (ARRUDA *et al.*, 2018), uma vez que os vértices são fragmentos maiores de texto (nesse caso uma quantidade limitada de parágrafos subsequentes) e as arestas conectam os vértices a partir de um critério de similaridade, no trabalho mencionado foi utilizada a similaridade de

cosseno. A Figura 2 ilustra a metodologia de construção de uma rede mesoscópica.

No trabalho (ARRUDA *et al.*, 2018), a rede complexa é construída a partir de um texto organizado, isto é, uma sequência de palavras dispostas em parágrafos. O texto organizado O pode ser representado por $O = (p_0, p_1, p_2 \dots)$, no qual cada parágrafo p_i é composto por uma sequência de palavras $p_i = (w_{i0}, w_{i1}, w_{i2} \dots)$. Assim, para a criação dos vértices define-se uma janela de tamanho Δ para contemplar Δ parágrafos adjacentes em O, $P_k^{\Delta} = (p_k, p_{k+1},\dots,p_{k+\Delta-1})$. Para definir as arestas, é necessário aplicar a similaridade de cosseno entre todos os pares de janela de textos P_k^{Δ} e manter apenas as conexões entre vértices que atingiram um valor de similaridade maior que um limite T, por fim resultando em uma rede não ponderada. Vale ressaltar que a sobreposição de parágrafos sucessivos na janela de textos para a formação dos vértices faz com que vértices sucessivos sempre estejam conectados, devido à semelhança do conteúdo compartilhado e sendo assim, a evolução temporal do texto é contemplada no modelo.

Um caso específico de redes mesoscópicas foi apresentado em (ARRUDA *et al.*, 2019), no qual o vértice é representado por um parágrafo único, sem ter uma sobreposição forçada entre fragmentos adjacentes. A vantagem de construir uma rede de abordagem baseada em parágrafo, é a possibilidade de analisar documentos nos quais a ordem das páginas e dos parágrafos são desconhecidos.

2.3 Medidas de redes

Há inúmeras maneiras de fazer a caracterização topológica de redes complexas, nesse trabalho serão abordadas as métricas tradicionais mencionadas na disciplina de redes complexas, além de métricas utilizadas em trabalhos anteriores.

2.3.1 Grau

O grau de um vértice aponta quantas ligações o mesmo possui, sendo o grau do vértice i representado na matriz de adjacência A da seguinte maneira:

$$k_i = \sum_{j=1}^{M} a_{ij} = \sum_{i=1}^{M} a_{ij} \tag{2.1}$$

Quanto maior o grau de um determinado vértice, indica o quão relevante é aquele vértice para a rede, assim, é possível capturar uma informação de centralidade. Essa medida pode ser aplicada em redes sociais para obter os usuários mais influentes, por exemplo. No caso deste trabalho, valores mais altos de grau indicam que o conteúdo dos vértices é mencionado em outros momentos da narrativa.

2.3.2 Diâmetro e excentricidade

Antes de falar sobre diâmetro e excentricidade, é importante entender o conceito de distância geodésica, que é o menor caminho entre dois vértices de uma rede. Então,

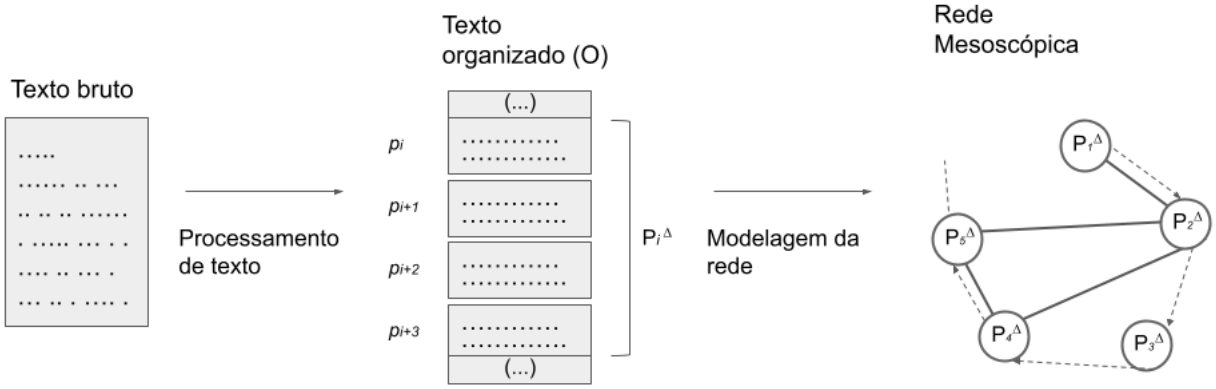

Figura 2 – Representação de texto para rede mesoscópica, em que primeiramente ocorre o processamento do texto bruto, sendo disposto em um texto organizado. A partir da definição de uma janela de parágrafos P_k^{Δ}, ocorre a modelagem da rede onde P_k^{Δ} são os vértices e as arestas se dão por meio da sequência e similaridade entre os parágrafos.

o diâmetro de uma rede é a maior distância geodésica entre quaisquer dois vértices (NEWMAN, 2003).

Já a excentricidade de um vértice v é a máxima distância entre v e qualquer outro vértice da rede G. O valor máximo de excentricidade resulta no diâmetro.

2.3.3 Centralidade de intermediação

Centralidade de intermediação (B) também conhecida como *betweenness*, indica quanto um vértice v está presente nos caminhos geodésicos da rede (FREEMAN, 1977), assim, um alto valor de centralidade de intermediação aponta que o vértice é bastante acessado na distância geodésica entre pares de vértices.

$$B(v) = \sum_{s,t \in G} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{2.2}$$

Onde s e t são vértices pertencentes a G, $\sigma(s,t)$ é a quantidade de caminhos geodésicos entre (s,t) e $\sigma(s,t|v)$ é a quantidade desses caminhos que passam por v.

2.3.4 Coeficiente de aglomeração

O coeficiente de aglomeração (CC) mede a densidade local de conexões entre vizinhos em um dado vértice e é dado pela relação entre o número de triângulos existentes dentre todas as possíveis conexões de um conjunto de três vértices (AMANCIO, 2015b).

$$CC = 3\sum_{k>j>i} a_{ij}a_{ik}a_{jk}\Big[\sum_{k>j>i} a_{ij}a_{ik} + a_{ji}a_{jk} + a_{ki}a_{kj}\Big]^{-1} \tag{2.3}$$

sendo que $0 <= CC_i <= 1$.

Essa métrica tem sido muito aplicada para verificar a presença de comunidades na rede e para diferenciar redes aleatórias de redes mundo pequeno (ALBERT; BARABASI, 2002).

2.3.5 Centralidade de proximidade

Centralidade de proximidade, ou *closeness*, mede o quão perto o vértice v está de todos os vertices da rede. Para isso, calcula-se o inverso da somatória da distância geodésica de v a $n-1$ vértices da rede e multiplica essa fração por $n-1$ para normalizar pela quantidade de caminhos mínimos possíveis (FREEMAN, 1979). Quanto maior o valor de centralidade de proximidade, mais próximo v está dos demais vértices portanto mais centralizado.

$$C(v) = \frac{n-1}{\sum_{u=1}^{n-1} d(u,v)} \tag{2.4}$$

onde $d(u,v)$ é o caminho mínimo entre v e u e n é o número de vértices do grafo.

2.3.6 Índice de correlação

O índice de correlação quantifica a similaridade entre duas regiões conectadas por uma aresta. A similaridade entre dois vértices ligados é calculada pela quantidade de conexões compartilhadas, isto é, pelo número de vizinhos comuns (ARRUDA *et al.*, 2018). Essa métrica pode ser escrita da seguinte forma:

$$\mu_{i,j} = \sum_{k\neq i,j} a_{ik}a_{jk}\Big[\sum_{k\neq j} a_{ik} + \sum_{k\neq i} a_{jk}\Big]^{-1} \tag{2.5}$$

onde a_{ij} compõe a matriz de adjacência representando a presença de aresta, quando existe uma conexão entre os vértices i e j, $a_{ij} = 1$, caso contrário, $a_{ij} = 0$.

2.3.7 Assinatura de recorrência

A assinatura de recorrência é uma medida desenvolvida para redes mesoscópicas apresentada em (SOUZA *et al.*, 2022). A medida consiste em capturar a ocorrência de alguma dobra na rede, ou seja, quando há alguma conexão por similaridade entre os nós, sem considerar as arestas de sequência. Além de registrar a referência cruzada, que pode ser

tanto no passado quanto no futuro da sequência de vértices, a medida também contabiliza a frequência em que a dobra acontece no enredo.

3 METODOLOGIA

Esta seção apresentará como a metodologia de transformar livros em redes mesoscópicas foi aplicada junto a um algoritmo de classificação a fim de identificar padrões entre as redes que pudessem indicar o sucesso de um livro.

Para tal, foi feita uma seleção de 218 livros em inglês disponíveis no Projeto Gutenberg, no qual foram submetidos a etapas de processamento de texto, modelagem da rede, extração de medidas e aplicação de um algoritmo supervisionado de aprendizado de máquina. Além disso, outras técnicas fora do campo de estudo de redes complexas foram abordadas com a finalidade de comparar resultados obtidos modelando o mesmo problema a partir de diferentes formas.

3.1 Pré-processamento

3.1.1 Base de dados

A base de dados utilizada neste trabalho provém de livros disponíveis no Projeto Gutenberg (PROJECT. . . ,), uma biblioteca online que oferece acesso gratuito a livros eletrônicos que não possuem proteção por direitos autorais nos Estados Unidos, ou seja, são de domínio público ou já tiveram seus direitos autorais expirados.

Na construção da base, foram selecionados 218 livros escritos na língua inglesa publicados no período de 1895 a 1923. Dentre as obras, 110 são consideradas *bestsellers* nos Estados Unidos, de acordo com a *Publishers Weekly*, uma revista semanal norte-americana voltada para o público do mercado de livros como editoras, livrarias e escritores.

3.1.2 Processamento de texto

O processamento do texto é necessário para manter apenas as informações mais importantes de cada parágrafo do livro, como está esquematizado na Figura 3. Primeiramente foi realizada uma limpeza dos dados através da remoção de marcadores de início de capítulo e caracteres como "_". Além disso, o texto foi padronizado em caixa baixa e definido que as palavras ficariam em seu estado canônico utilizando a técnica de lematização, procedimentos importantes para ter um melhor resultado ao aplicar a similaridade de cosseno posteriormente.

Em seguida, com o intuito de evitar palavras diferentes se referindo à mesma entidade e assim reduzir ruído ao analisar como as palavras se relacionam na frase, é executada a resolução de anáforas. Dessa forma, um exemplo do resultado obtido com a resolução de anáforas é a substituição de um pronome que se refere a uma menção anterior,

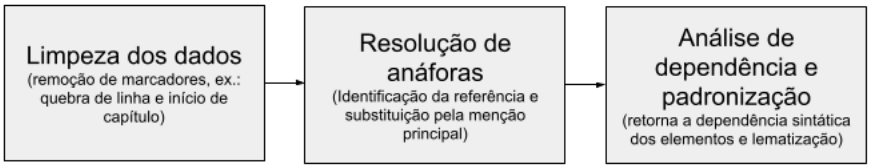

Figura 3 – Diagrama de execução do processamento de texto. Etapas realizadas para cada livro.

por essa menção principal: *"Keith nodded. "You won,"he said."* para *"Keith nodded. "You won,"Keith said."* do livro *The River's End*.

Após isso, realizou-se a análise de dependência sintática na qual sujeito, verbo e objeto direto são identificados e mantidos a fim de reter no parágrafo apenas as relações com maior relevância de conteúdo para o enredo. A aplicação da lematização é feita enquanto captura-se o objeto direto e o predicado nominal. As duas últimas etapas, resolução de anáforas e análise de dependência sintática, foram efetuadas através da implementação de bibliotecas e modelos pré-treinados de processamento de linguagem natural (NLP), como o pacote *spaCy* e sua extensão *NeuralCoref* (MANNING *et al.*, 2014).

Vale ressaltar que o processamento de texto descrito foi utilizado apenas para a abordagem de rede mesoscópica, a preparação dos dados para as outras abordagens estará na seção 3.2.

3.1.3 Modelagem da rede

O objetivo deste tópico é a partir do texto organizado gerado (O) por meio do processamento de texto, criar uma rede mesoscópica em que os vértices serão um conjunto de parágrafos subsequentes, interligados por meio da similaridade de cosseno entre eles. Essas etapas estão ilustradas na Figura 4. Importante ressaltar que cada livro dará origem a um grafo resultante da modelagem da rede.

Deste modo, o texto organizado O é formado por uma sequência de parágrafos $O = (p_0, p_1, p_2 \dots)$ que por sua vez possui uma série de palavras $p_i = (w_{i0}, w_{i1}, w_{i2} \dots)$. Para formar os vértices da rede mesoscópica é definido um tamanho de janela $(p_k, p_{k+1},\dots,p_{k+\Delta-1})$ que percorre a sequência de parágrafos e atribui essa composição ao vértice P_k^{Δ}. Logo, os

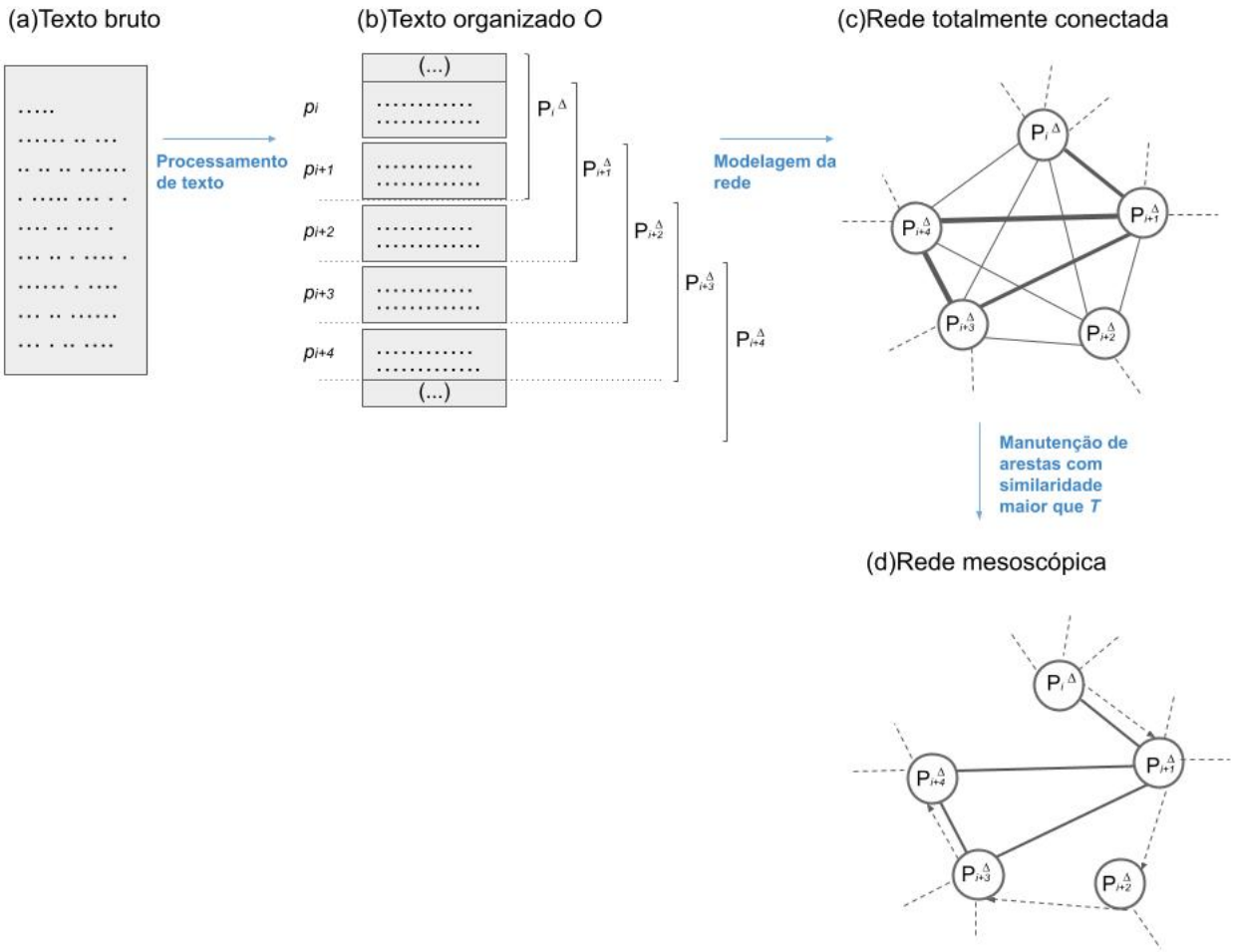

Figura 4 – Etapas da transformação do texto em rede complexa. O processamento de texto ocorre e o texto bruto (a) se torna um texto organizado O (b). Após isso, acontece a etapa de modelagem da rede, onde cada janela de parágrafos (P_k^{Δ}) de O se torna um vértice de uma rede totalmente conectada (c), em que as conexões são formadas pela similaridade de cosseno. Por fim, a rede mesoscópica é formada permanecendo apenas as arestas de similaridade relevante (linhas contínuas) e as arestas entre parágrafos consecutivos (arestas sequenciais representadas pelas setas pontilhadas).

vértices consecutivos terão parágrafos em comum.

Para obter o peso das arestas, foi extraída a medida *TF-IDF*, no qual multiplica a frequência de um termo t em um dado documento d com o inverso da frequência do documento (*idf*):

$$TF - IDF(t, d) = tf(t, d) * idf(t) \tag{3.1}$$

Sendo:

$$idf(t) = log\frac{1+n}{1+df(t)} + 1 \tag{3.2}$$

onde n = quantidade total de documentos e $df(t)$ é a frequência do documento contendo o termo t.

Nesse caso, o número total de documentos representa a quantidade de vértices, o documento d é um dos vértices composto pela janela de Δ parágrafos e o termo t corresponde a cada palavra w do parágrafo. Dado isso, cada vértice configura um vetor de representação *bag of words* com os valores de TF-IDF para w contidos. Por fim, é calculado uma matriz de similaridade S contendo a similaridade de cosseno entre os vértices onde:

$$similaridade = \cos(\Theta) = \frac{\mathbf{A}\ \mathbf{B}}{\|\mathbf{A}\|\,\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_{i^2}}} \tag{3.3}$$

Com os vértices ordenados e os pesos calculados em S, é formada uma rede totalmente conectada ponderando as ligações em que a posição dos vértices tenha uma diferença $> 2 * \Delta$, garantindo dessa forma que vértices com parágrafos em comum não traga algum tipo de viés para a rede.

A rede mesoscópica finalmente é alcançada quando as conexões mais fracas são removidas de modo que o grau médio da rede atinja um limite definido T, última etapa da Figura 5. Ademais, a rede mesoscópica de cada livro terá não somente a ligação determinada pela similaridade, mas também uma ligação denominada "sequencial" que conecta vértices subsequentes que assegurará a perspectiva temporal da narrativa. Dessa forma, cada nó pode conter até dois tipos de arestas conectando os demais nós da rede, as arestas de similaridade e sequencial.

3.1.4 Extração de medidas das redes

Uma vez que a rede mesoscópica está criada, é possível extrair medidas referentes a topologia da rede. Utilizando a biblioteca *NetworkX*, inicialmente foram calculadas medidas de distância de centralidade para cada livro, sendo elas: diâmetro, excentricidade, coeficiente de intermediação (*betweenness*), coeficiente de aglomeração (*clustering*), centralidade de proximidade (*closeness*). Já para o experimento 2, incluiu-se outras duas medidas específicas de redes mesoscópicas, que são o índice de correlação e assinatura de recorrência.

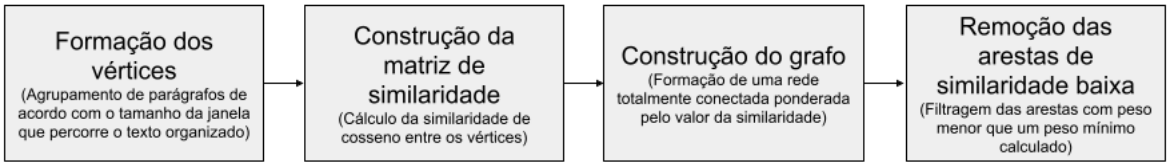

Figura 5 – Diagrama de execução da transformação do texto em rede complexa.

Com exceção do diâmetro e da assinatura de recorrência, as demais medidas retornam um valor por vértice ou aresta. Com o intuito de extrair um valor global da rede para cada métrica, foi realizada a média aritmética de todos os vértices ou arestas da rede. Assim, a partir do coeficiente de aglomeração de cada nó obtém-se o coeficiente de aglomeração médio da rede, por exemplo. Por fim, o pré-processamento é concluído e tem como resultado uma base de dados na qual cada linha representa um livro com suas respectivas medidas calculadas e a marcação de sucesso.

3.2 Abordagens sem considerar redes complexas

Em mineração de texto existem várias formas de organizar os dados de forma estruturada, além das técnicas que envolvem redes complexas. Neste trabalho, duas outras formas de processar texto foram aplicadas para gerar variáveis utilizadas nos classificadores. são elas: *TF-IDF* e *embedding*. O objetivo é comparar a performance dos modelos em diferentes abordagens para a predição de livros de sucesso.

3.2.1 *TF-IDF*

TF-IDF é uma medida estatística utilizada em pré-processamento de textos capaz de avaliar a importância de um termo t em relação à coleção de documentos. Dessa forma, uma palavra presente em muitos documentos (alto valor de df) sugere que é um termo muito comum e quando é calculada a frequência inversa (idf) resulta em uma baixa importância. Essa medida é obtida a partir das equações 3.1 e 3.2 mencionadas no item 3.1.3.

O intuito de empregar o *TF-IDF* é entender se os termos utilizados na escrita interferem no sucesso de vendas de um livro. Para tal, cada livro teve seu texto transformado em uma lista de sub-sequências (*tokenizer*), realizou-se a remoção de *stopwords* e em seguida a redução das palavras ao seu radical (*stemming*), ambos utilizando pacotes da biblioteca *nltk*. Por fim, foi gerada a matriz *TF-IDF* de tamanho $(218, 70641)$, em que a quantidade de linhas refere-se ao número de documentos, que nesse caso equivale aos livros contidos na base de dados e as colunas são os termos únicos presentes nos documentos.

3.2.2 *Embedding*

O *TF-IDF* por ser uma medida simples, tem dificuldade em identificar a similaridade entre textos formados de palavras diferentes, porém sinônimos. Dada essa limitação, foi abordada a técnica de representação distribuída que utiliza redes neurais, também conhecida como *embeddings*.

Palavras, parágrafos ou documentos podem ser representados como um vetor. Assim, para gerar uma representação de cada livro foi empregada a técnica *Doc2Vec* baseada em (LE; MIKOLOV, 2014), na qual cada livro foi representado por um vetor de tamanho 64, parâmetro escolhido para gerar o vetor. *Doc2Vec* foi aplicado utilizando a biblioteca *Gensim*.

3.3 Critérios de avaliação

Este projeto visa processar o texto de livros que obtiveram ou não sucesso de três formas distintas: redes mesoscópicas, *TF-IDF* e *Doc2Vec*. Após essa etapa foram testados alguns algoritmos supervisionados a fim de predizer os livros de sucesso.

Para avaliar o resultado dos classificadores, a base de dados foi dividida em dados de treino e teste, sendo que a amostra de teste corresponde a 25% da base total. Nos dados de teste foi aplicada a métrica de classificação acurácia, que compara a variável resposta com o resultado predito do modelo.

Um conceito importante para entender o cálculo da métrica de avaliação é a matriz de confusão (Figura 6), onde o eixo x representa as classes preditas pelo modelo e o eixo y as classes reais. Em cada quadrante está a quantidade de observações para a combinação da classe predita e real, sendo a diagonal principal os casos em que o modelo acertou, ou seja, no primeiro quadrante são os verdadeiros negativos (VN) e no quarto quadrante os verdadeiros positivos (VP). Já no segundo quadrante estão os falsos positivos (FP) e no terceiro quadrante os falsos negativos (FN).

O cálculo da acurácia pode ser resumido pela soma da quantidade de casos em que o modelo acertou na predição sobre todos os casos.

$$acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.4}$$

Neste caso a acurácia consegue expressar o desempenho dos modelos de forma adequada, pois na base de dados as classes estão balanceadas. Caso contrário a escolha de outra métrica, como o *f1-score* por exemplo, seria mais apropriado.

Além disso, para evitar que as amostras de treino e teste tragam alguma interpretação enviesada para a avaliação dos algoritmos de classificação, também foi aplicada a técnica de validação cruzada, na qual é possível treinar o modelo com diferentes combinações da base de treino e para cada combinação avaliar a manutenção do desempenho.

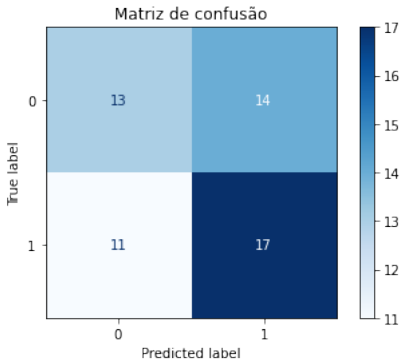

Figura 6 – Exemplo de matriz de confusão, em que o eixo x é o valor predito e o eixo y é o valor real. Neste caso o valor da acurácia é 0.55.

Para isso, a base de dados é dividida em k repartições menores, na qual a cada rodada o modelo utiliza $k-1$ repartições para o aprendizado e a predição é aplicada na repartição que ficou de fora do treino. Neste trabalho a validação cruzada foi feita considerando 4 repartições e analisando a média e desvio padrão das métricas.

4 RESULTADOS

Nesta seção os resultados dos quatro experimentos realizados serão apresentados. Nos dois primeiros experimentos, foi aplicado aos livros a abordagem de redes mesoscópicas, divergindo entre eles apenas nos parâmetros adotados para montar as redes e as medidas extraídas. Já no experimento 3 a técnica empregada para tratar o texto foi *TF-IDF* e no experimento 4 foi utilizado *Doc2Vec*. Em todos os experimentos após o processamento do texto foram treinados os seguintes classificadores: *k-vizinhos mais próximos (k-NN), Random Forest, Suppport Vector Machine (SVM) e Multilayer Perceptron (MLP)*, escolhidos de forma empírica dentre alguns algortimos de aprendizado de máquina testados.

4.1 Experimento 1

Neste experimento foi gerada uma rede mesoscópica a partir de cada livro como descrito no item 3.1.3 considerando o tamanho da janela $\Delta = 4$ e o grau médio parametrizado em 3, definidos de maneira empírica. Vale ressaltar que as medidas extraídas das redes foram: diâmetro, excentricidade, centralidade de intermediação (*betweenness*), coeficiente de aglomeração (*clustering*) e centralidade de proximidade (*closeness*), sendo considerado o valor da média para as últimas quatro variáveis.

Dois livros foram escolhidos com o intuito de exemplificar como as redes se configuraram e apresentar as medidas extraídas. O livro *If winter comes* da classe de sucesso e o livro *The infidel* da classe dos livros que não se tornaram *bestseller* tiveram suas redes projetadas na Figura 7. Na tabela 1 é possível notar que apesar da quantidade de vértices e arestas de ambos os livros terem grandezas parecidas, o diâmetro da rede do livro *If winter comes* (sucesso=1) é o dobro do diâmetro de *The infidel* (sucesso=0). Assim, redes de livros com quantidade de vértices de mesmo patamar ainda podem ter topologias diferentes, pois também dependem das arestas de similaridade que são formadas, afetando os caminhos mínimos e centralidade dos vértices.

4.1.1 Extração de padrões

Nessa etapa foi analisado se as variáveis explicativas geradas a partir das medidas da rede são capazes de explicar o sucesso dos livros e a relação entre essas variáveis.

4.1.1.1 Análise gráfica e correlação

Inicialmente, foi feita uma análise exploratória dos dados com o objetivo de entender tanto a distribuição das observações em relação a classe de sucesso, quanto para avaliar a correlação entre as variáveis.

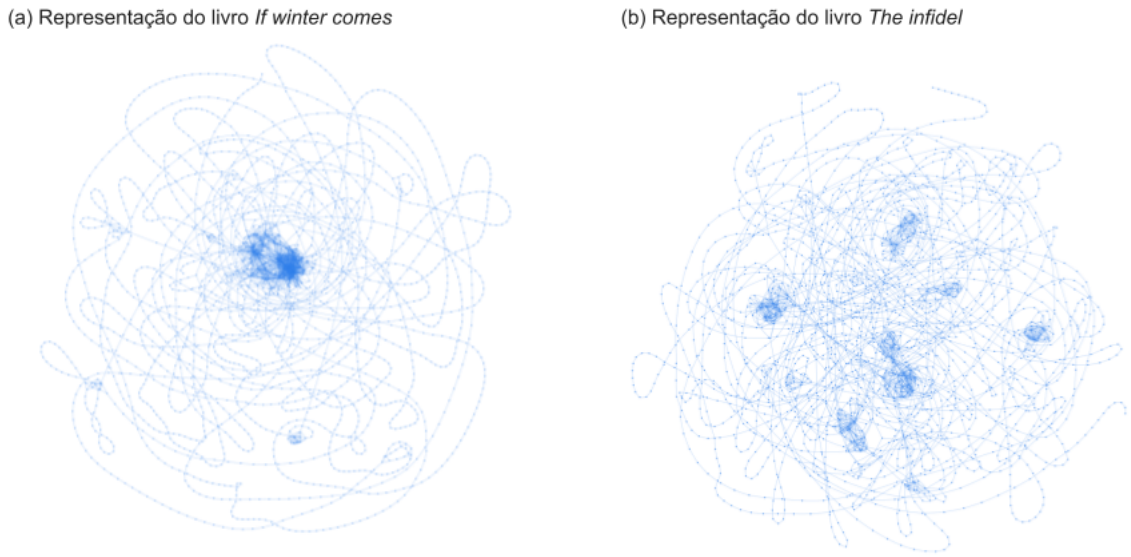

Figura 7 – Visualização das redes complexas calculadas com parâmetro $\Delta = 4$ e grau médio = 3. (a) Livro *If winter comes*, pertencente a classe de sucesso e (b) Livro *The infidel*, pertencente a classe de demais livros.

Tabela 1 – Informações e medidas provenientes da rede mesoscópica de dois livros no experimento 1.

Livro	*If winter comes*	*The infidel*
Δ	4	4
Grau médio	3	3
Quantidade vértices	2372	2383
Quantidade arestas	3559	3575
Quantidade arestas de sequência	2371	2382
Quantidade arestas de similaridade	1188	1193
Diâmetro	289	143
Excentricidade	188.473	101.370
Betweenness	0.030	0.014
Clustering	0.047	0.078
Closeness	0.017	0.033
Marcação sucesso (binário)	1	0

Uma primeira análise gráfica foi feita, projetando pares de variáveis em uma matriz, onde a diagonal é composta por histogramas e o restante por gráficos de dispersão (Figura 8), sendo que as observações dos livros de sucesso estão representadas pela cor laranja e os demais livros na cor azul. Nota-se que não há uma divergência significativa da distribuição dos dados nos histogramas em relação às classes, uma vez que as barras estão sobrepostas.

Analisando os gráficos de dispersão da Figura 8, também não se percebe formação de grupos apartados de acordo com a classe. No entanto, em certas combinações de eixos é possível observar uma ordenação dos dados, propondo uma correlação entre as variáveis, como por exemplo no gráfico diâmetro versus excentricidade.

A hipótese de que existem variáveis correlacionadas se confirma ao aplicar o coeficiente de correlação de *Pearson*, apresentada na Tabela 2. O maior coeficiente de correlação de *Pearson* ocorre para as variáveis de diâmetro e excentricidade, com o valor de 0.99, indicando uma alta correlação linear positiva, o que já era esperado, visto que as duas medidas levam em consideração o conceito de caminhos mínimos. Há também correlações negativas, entre centralidade de proximidade e diâmetro e centralidade de proximidade e excentricidade, ambas com um coeficiente de -0.73.

Tabela 2 – Coeficiente de correlação de *Pearson* entre as variáveis do experimento 1.

Variáveis	closeness	clustering	betweenness	excentricidade	diâmetro
closeness	1.00	0.52	-0.41	-0.73	-0.73
clustering	0.52	1.00	0.04	-0.47	-0.50
betweenness	-0.41	0.04	1.00	0.53	0.49
excentricidade	-0.73	-0.47	0.53	1.00	0.99
diâmetro	-0.73	-0.50	0.49	0.99	1.00

4.1.1.2 Análise de componentes principais (PCA)

Posteriormente, realizou-se uma análise de componentes principais com a finalidade de avaliar a importância de cada componente para explicar a variável de classificação de sucesso e possivelmente uma redução na dimensionalidade da base de dados. Para isso, foi utilizado o módulo *decomposition* da biblioteca *sklearn*, além de realizar a normalização dos dados. Assim, ao rodar a análise para cinco componentes e verificar a variância explicada por cada componente principal, tem-se que as duas primeiras componentes principais juntas contêm 86% da informação para explicar a variável resposta. Na Figura 9 está disposta a variância de cada componente principal.

Apesar das componentes principais 1 e 2 obterem a maior parte da informação, ao plotar um gráfico de dispersão com essas duas componentes identificando as classes (Figura 10), não há uma separação clara entre os livros classificados como sucesso dos demais.

Ao calcular a importância de cada variável para as duas primeiras componentes principais em Tabela 3, observa-se em $PC1$ que a variável com maior importância é o diâmetro, já em $PC2$, *betweenness* tem o maior peso. Ambas as medidas tem como base o conceito de distância geodésica.

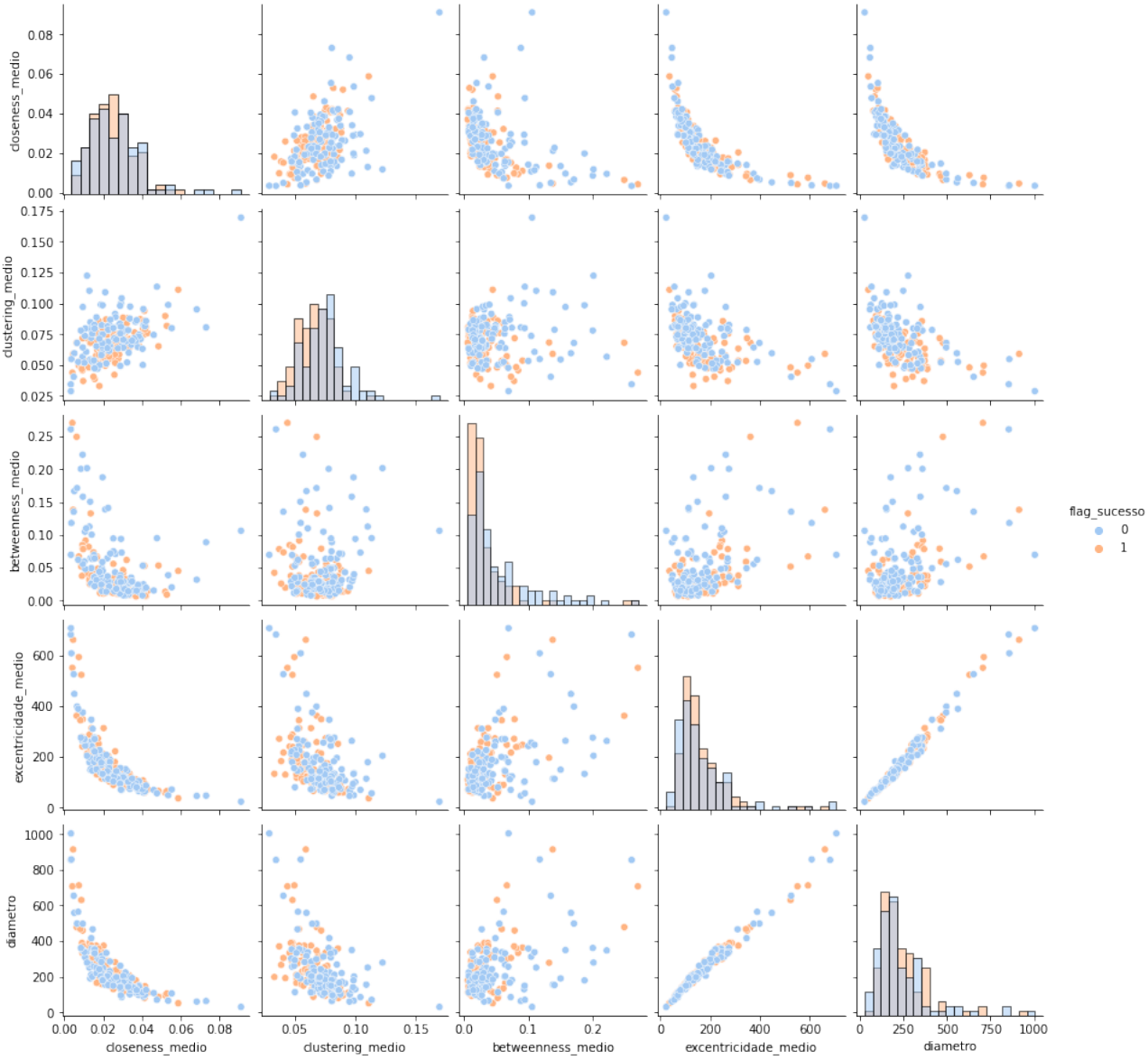

Figura 8 – Gráficos de dispersão de pares de medidas e histogramas no experimento 1.

4.1.2 Aplicação dos classificadores

Por fim, a partir das medidas geradas com base em cada rede e a marcação das classes, foram treinados quatro modelos supervisionados de classificação diferentes com a variação utilizando ou não as componentes do PCA. Na divisão da base em treino e teste, foi usada uma proporção de 25% para teste. Para o aprendizado do modelo com os dados resultantes do PCA, o número de componentes parametrizados foi de 3.

Na Tabela 4, o melhor resultado de acurácia ocorreu para o classificador *MLP* utilizando os dados do PCA (0.80), no entanto quando se compara com o valor da acurácia média obtida através da validação cruzada (0.63), existe uma diferença de 0.17, a maior dentre os cenários apresentados, além dessa diferença superar o desvio padrão. Analisando

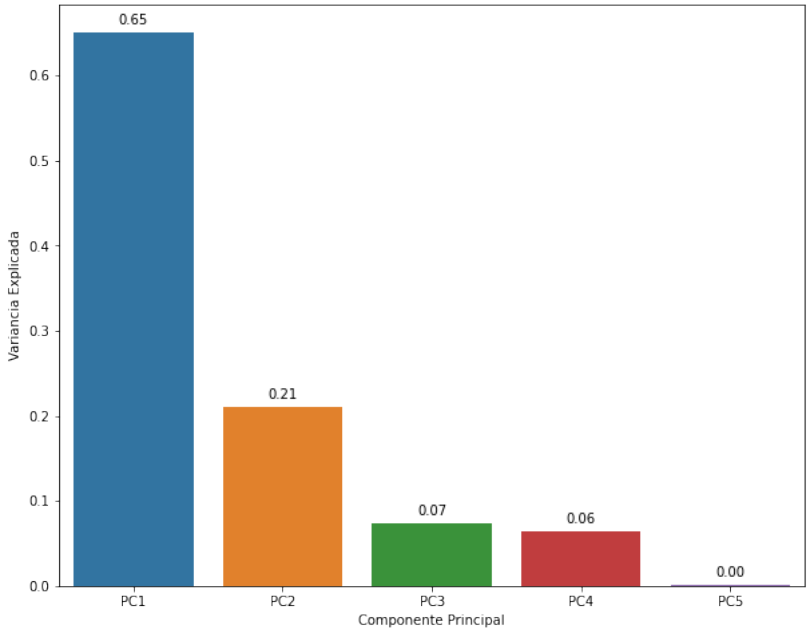

Figura 9 – Variância explicada de acordo com cada componente principal do experimento 1.

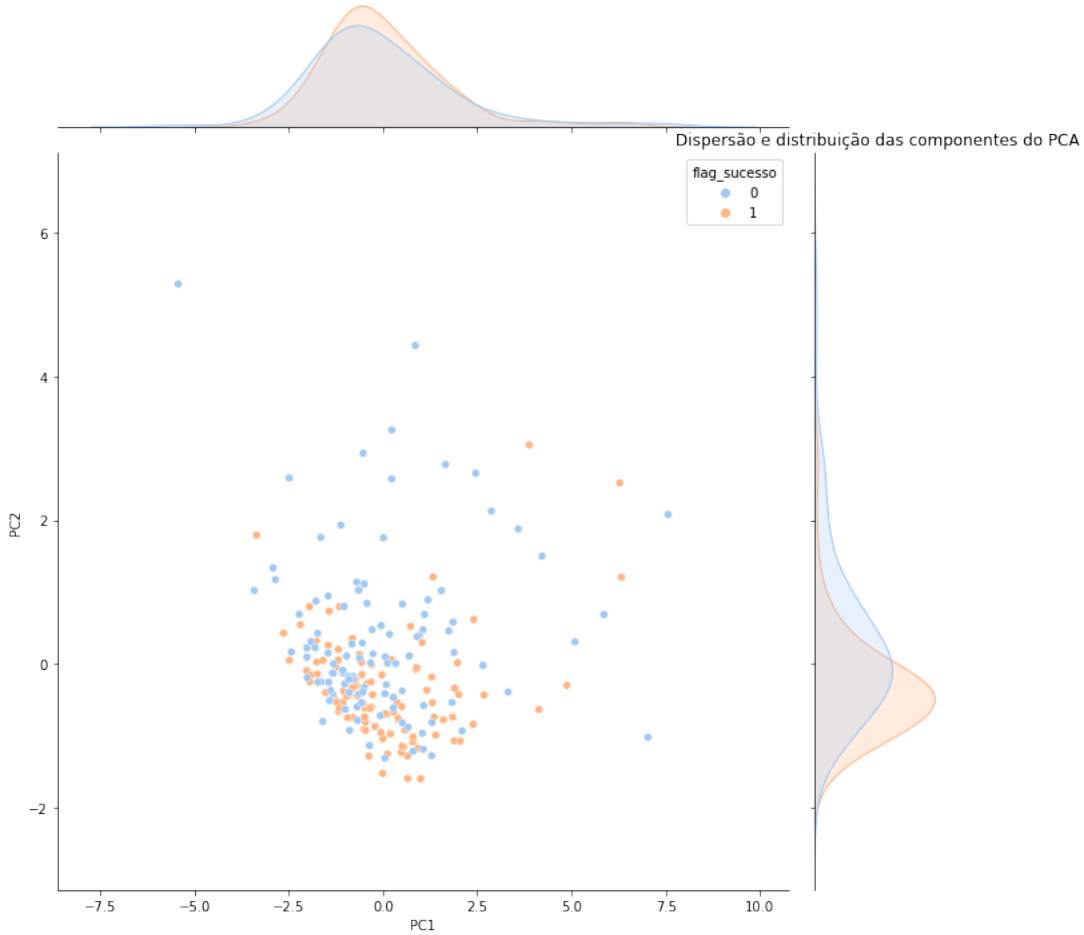

Figura 10 – Gráfico de dispersão e histograma para visualização das componentes principais 1 e 2 por classe do experimento 1.

Tabela 3 – Ordenação das variáveis de acordo com a importância em cada componente principal do experimento 1.

Componente principal	Variáveis	Importância normalizada
PC1	diâmetro	24.23%
	excentricidade	24.21%
	closeness	21.79%
	clustering	15.23%
	betweenness	14.54%
PC2	betweenness	45.33%
	clustering	43.79%
	closeness	5.56%
	excentricidade	4.24%
	diâmetro	1.09%

esses valores, há um possível caso de *overfitting*, no qual em um conjunto de dados específicos o modelo consegue discriminar muito bem as classes, porém em outros dados não consegue generalizar.

Por consequência, considerando tanto o valor da maior acurácia média quanto o menor desvio padrão, o classificador com melhor desempenho foi o *SVM* utilizando as componentes do PCA, embora os demais modelos possuírem patamares próximos de grandeza da acurácia média.

Tabela 4 – Avaliação dos resultados obtidos por cada classificador no experimento 1.

Classificador	PCA	Acurácia	Acurácia média	Desvio padrão
k-NN	Não	0.64	0.57	0.11
	Sim	0.62	0.58	0.06
Random Forest	Não	0.64	0.56	0.06
	Sim	0.67	0.59	0.06
SVM	Não	0.73	0.60	0.04
	Sim	0.73	0.62	0.05
MLP	Não	0.58	0.57	0.03
	Sim	0.80	0.63	0.08

4.2 Experimento 2

As etapas realizadas neste experimento são as mesmas executadas no experimento 1, apenas modificando o parâmetro de grau médio adotado, que para esse caso foi igual a 5. Além disso, duas medidas de rede mesoscópicas foram adicionadas às variáveis, índice de correlação e assinatura de recorrência.

É possível reparar um aumento na quantidade de arestas tanto na Figura 11 quanto na Tabela 5 em decorrência do aumento do parâmetro do grau médio. Na Tabela 6 pode-se observar que a mudança do grau médio não só afetou a quantidade de arestas, como

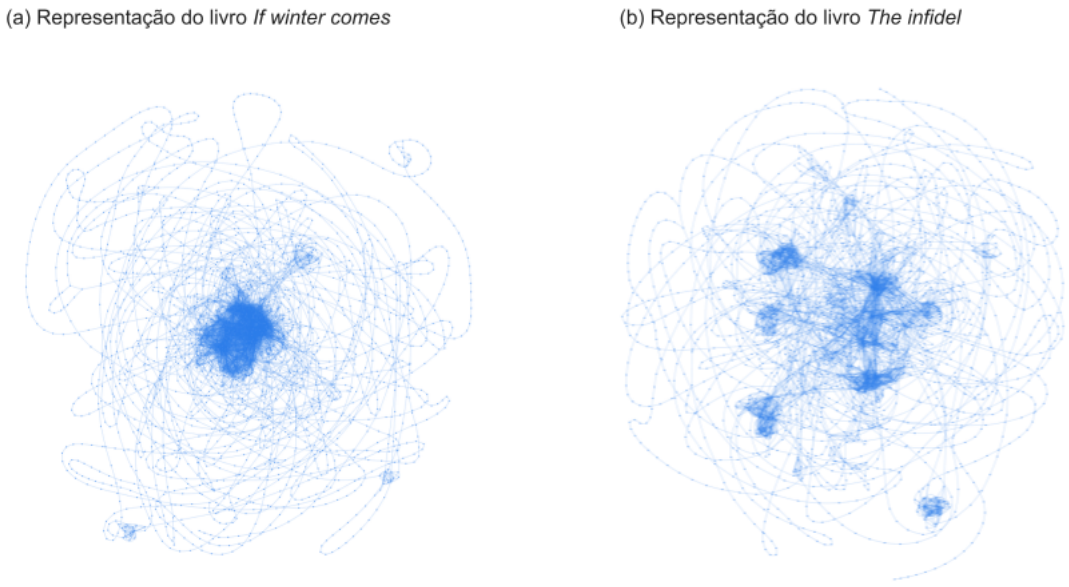

Figura 11 – Visualização das redes complexas calculadas com parâmetro $\Delta = 4$ e grau médio = 5. (a) Livro *If winter comes*, pertencente a classe de sucesso e (b) Livro *The infidel*, pertencente a classe de demais livros.

também reduziu as médias das medidas baseadas em caminhos geodésicos, com exceção de *closeness*. Além disso, houve um aumento na métrica baseadas em densidade, *clustering*.

Tabela 5 – Informações e medidas provenientes da rede mesoscópica de dois livros no experimento 2.

Livro	*If winter comes*	*The infidel*
Δ	4	4
Grau médio	5	5
Quantidade vértices	2372	2383
Quantidade arestas	5931	5958
Quantidade arestas de sequência	2371	2382
Quantidade arestas de similaridade	3560	3576
Diâmetro	196	84
Excentricidade	136.244	57.646
Betweenness	0.015	0.006
Clustering	0.101	0.147
Closeness	0.033	0.074
Índice de correlação	0.097	2.644
Assinatura de recorrência	4.116	0.106
Marcação sucesso (binário)	1	0

Tabela 6 – Comparação das médias das medidas extraídas no experimento 1 e 2.

Medidas	Média Experimento 1	Média Experimento 2
diâmetro	238.394	118.101
excentricidade	170.070	81.199
betweenness	0.045	0.017
clustering	0.070	0.136
closeness	0.025	0.060
índice de correlação	-	0.115
assinatura de recorrência	-	3.112

4.2.1 Extração de padrões

4.2.1.1 Análise gráfica e correlação

As relações entre as variáveis apontadas no experimento 1, se mantêm para o experimento 2 como pode ser observado na Figura 12. Para as variáveis adicionadas, a assinatura de recorrência possui correlação positiva com o diâmetro (0.60) e com excentricidade (0.57), além de ter correlação negativa com *closeness* (-0.64) e com *clustering* (-0.89).

4.2.1.2 Análise de componentes principais (PCA)

A análise de componentes principais para o experimento 2 trouxe um resultado de variância explicada em patamares semelhantes ao experimento 1, em que as duas primeiras componentes principais conseguem reter 80% da informação para explicar o problema.

Na Figura 13, mesmo com a distribuição das classes terem comportamentos distintos, no gráfico de dispersão ainda não há uma partição linear que consiga agrupar as classes. No entanto, comparado ao gráfico gerado no experimento 1 (Figura 10), é possível notar uma separação mais evidente entre a classe de sucesso e a classe do restante dos livros.

Na importância de cada variável para as componentes (Tabela 7), assim como ocorreu no experimento 1, em $PC1$ o diâmetro continua tendo o maior valor. Já em $PC2$ o índice de correlação recebeu maior relevância.

4.2.2 Aplicação dos classificadores

Para rodar os classificadores na base de dados do experimento 2, foram utilizadas as mesmas configurações do experimento 1. A Tabela 8 possui os valores de acurácia, acurácia média e desvio padrão, os dois últimos decorrentes da validação cruzada. É possível observar que a acurácia média desse experimento comparada às do experimento 1 conseguiu atingir valores maiores, indicando uma contribuição positiva das alterações realizadas na construção das redes e na inclusão de mais duas medidas como variáveis explicativas.

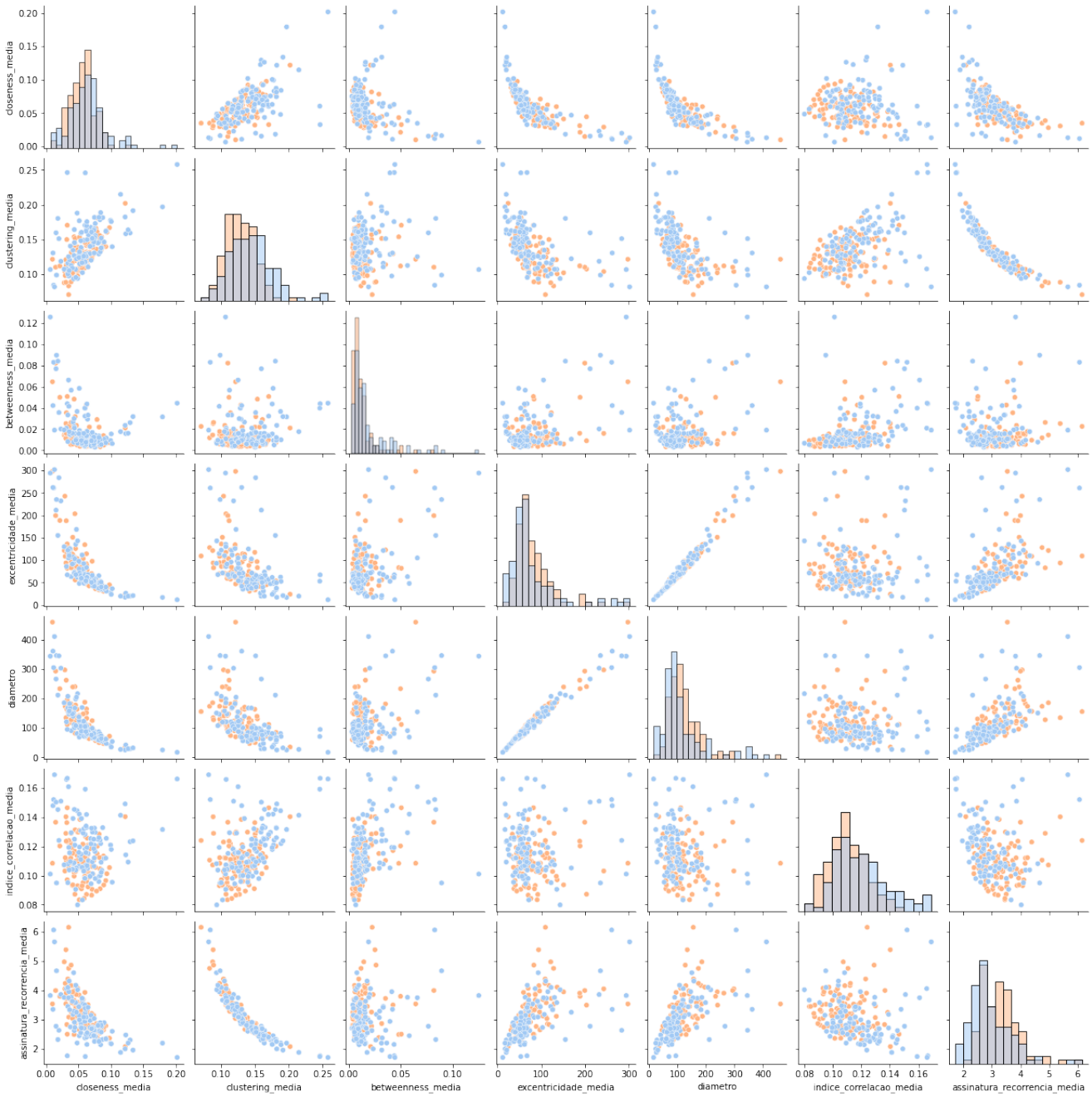

Figura 12 – Gráficos de dispersão de pares de medidas e histogramas do experimento 2.

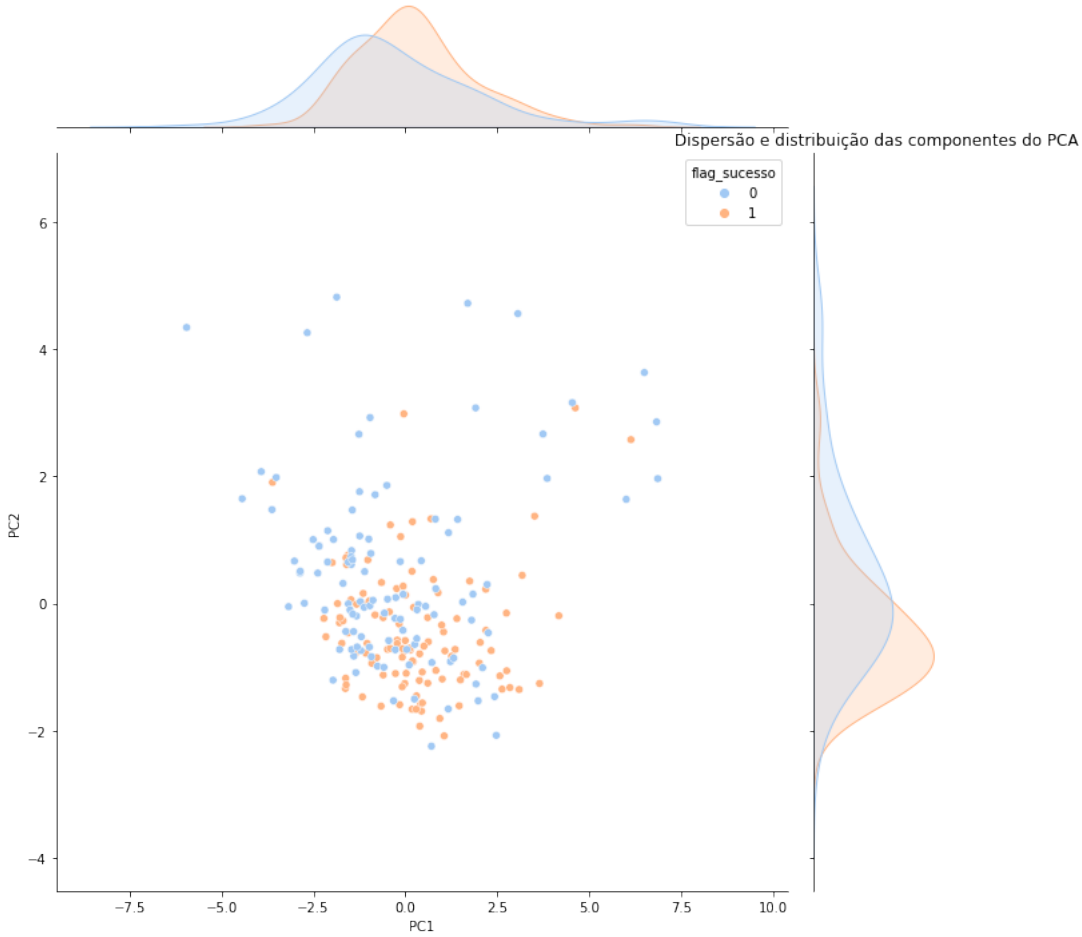

Figura 13 – Gráfico de dispersão e histograma para visualização das componentes principais 1 e 2 por classe do experimento 2.

Assim, ao analisar os resultados da Tabela 8 considerando a melhor combinação de acurácia média e desvio padrão, o classificador *MLP* sem utilizar as variáveis do PCA se destaca pela sua performance com uma acurácia média de 0.67.

4.3 Experimento 3

Neste experimento foi aplicada a técnica *TF-IDF* que resultou em uma matriz de tamanho $(218, 70641)$, dessa forma, cada palavra se tornou uma variável na qual armazena o valor de *TF-IDF* referente aos livros.

No experimento 3, foi utilizado o PCA para analisar as duas principais componentes da base de desenvolvimento criada a partir da matriz resultante do *TF-IDF*. No entanto, as duas primeiras componentes contêm apenas 13% da informação, não tornando vantajoso o uso das componentes do PCA para o treino dos algoritmos de classificação.

Na Tabela 9 estão os valores de acurácia para os classificadores testados, no qual o *MLP* apresenta resultados promissores, tanto na acurácia (0.71) para uma única amostra de teste, quanto na acurácia média (0.70) da validação cruzada.

Tabela 7 – Ordenação das variáveis de acordo com a importância em cada componente principal do experimento 2.

Componente principal	Variáveis	Importância normalizada
PC1	diâmetro	19.24%
	excentricidade	18.90%
	closeness	18.18%
	assinatura de recorrência	17.43%
	clustering	15.98%
	betweenness	8.67%
	índice de correlação	1.60%
PC2	índice de correlação	27.75%
	betweenness	24.18%
	clustering	19.67%
	assinatura de recorrência	10.70%
	excentricidade	8.91%
	diâmetro	6.85%
	closeness	1.95%

Tabela 8 – Avaliação dos resultados obtidos por cada classificador no experimento 2.

Classificador	PCA	Acurácia	Acurácia média	Desvio padrão
k-NN	Não	0.55	0.61	0.06
	Sim	0.64	0.61	0.04
Random Forest	Não	0.62	0.61	0.03
	Sim	0.73	0.58	0.04
SVM	Não	0.69	0.66	0.07
	Sim	0.69	0.62	0.05
MLP	Não	0.65	0.67	0.07
	Sim	0.69	0.63	0.04

Tabela 9 – Avaliação dos resultados obtidos por cada classificador no experimento 3.

Classificador	Acurácia	Acurácia média	Desvio padrão
k-NN	0.64	0.58	0.06
Random Forest	0.71	0.66	0.06
SVM	0.64	0.62	0.04
MLP	0.71	0.70	0.04

4.4 Experimento 4

No experimento 4, o texto dos livros foi processado utilizando o *Doc2Vec*, uma técnica que transforma um documento em um vetor usando *embeddings*. Foi padronizado para esse experimento que o tamanho do vetor seria 64, dessa forma, cada livro teria 64 valores que os representasse.

Assim como no experimento 3, para testar os classificadores não foi considerado as componentes obtidas através do PCA. Na Tabela 10, é possível observar que o *SVM* foi o

modelo que obteve melhor desempenho devido ao maior valor de acurácia média (0.69) e o baixo desvio padrão (0.03).

Tabela 10 – Avaliação dos resultados obtidos por cada classificador no experimento 4.

Classificador	Acurácia	Acurácia média	Desvio padrão
k-NN	0.65	0.58	0.06
Random Forest	0.58	0.59	0.06
SVM	0.62	0.69	0.03
MLP	0.55	0.58	0.03

5 CONCLUSÃO

A partir do desafio que as editoras possuem de avaliar se devem ou não publicar um determinado livro visando o sucesso de vendas e aliado ao avanço nas pesquisas tanto na área de mineração de textos quanto no campo de redes complexas para modelar textos, foi possível explorar se padrões intrínsecos aos livros contribuem para um resultado de sucesso.

A abordagem escolhida para transformar textos em redes complexas foi através de uma perspectiva mesoscópica, com a finalidade de analisar a influência que o contexto e o desdobramento temporal da história possuem em relação aos livros que entraram ou não na lista de *bestsellers*. Para isso, foram considerados uma seleção de 218 livros disponíveis no Projeto Gutenberg, no qual foram submetidos a etapas de processamento de texto, modelagem da rede, extração de medidas e aplicação de algoritmos supervisionados de aprendizado de máquina.

No experimento 1, construindo as redes mesoscópicas com parâmetros $\Delta = 4$ para a janela de parágrafos, grau médio de 3 e utilizando medidas clássicas de redes complexas como variáveis explicativas, foi alcançado uma acurácia média de 0.62 por meio do classificador *SVM* treinado com as componentes derivadas do PCA. Com a análise das componentes principais, foi possível verificar que o diâmetro e o *betweenness* médio são as medidas mais importantes para classificar os livros de sucesso, ambas levam em consideração a distância geodésica no cálculo. Assim, a forma como as conexões ocorrem ao longo da história e as possibilidades de caminhos que elas criam têm algum impacto na separação dos livros considerados *bestsellers*.

No experimento 2, os parâmetros escolhidos para gerar as redes mesoscópicas foram $\Delta = 4$ para a janela de parágrafos e grau médio de 5. Além disso, duas medidas específicas de redes mesoscópicas foram adicionadas às variáveis explicativas, sendo elas o índice de correlação e a assinatura de recorrência. Nesse cenário, foi observado o aumento da quantidade média de arestas por similaridade e redução no diâmetro médio das redes, o que já era esperado com o incremento no grau médio definido. Por fim, foi possível atingir uma acurácia média de 0.67 com o classificador *MLP*. Ademais, as medidas com maiores relevâncias no PCA foram o diâmetro e o índice de correlação médio. Dessa maneira, tanto os caminhos mínimos quanto as conexões em comum entre os vértices são fatores que conseguem contribuir na discriminação das classes.

Nos experimentos 3 e 4 foram aplicadas técnicas não pertencentes a esfera de redes complexas para processar o texto dos livros, *TF-IDF* e Doc2Vec respectivamente. O objetivo era avaliar o desempenho das redes mesoscópicas em relação a abordagens

mais tradicionais em processamento de linguagem natural (NLP). Para o experimento 3 obteve-se uma acurácia média de 0.70 com o classificador *MLP*, já para o experimento 4 o algoritmo *SVM* alcançou um resultado de 0.69. Apesar do experimento 2 chegar num valor de acurácia menor do que os experimentos 3 e 4, não há uma diferença discrepante entre os resultados, o que abre portas para mais possibilidades serem testadas com redes mesoscópicas.

Por conta do problema abordado neste trabalho ser complexo e também depender de muitos fatores que são externos ao texto dos livros, apenas as informações extraídas das redes nos experimentos 1 e 2 ou o processamento de texto utilizando técnicas de NLP nos experimentos 3 e 4 não foram suficientes para determinar com precisão o sucesso dos livros. Entratanto, os modelos supervisionados treinados a partir de redes mesoscópicas podem se tornar uma ferramenta capaz de extrair padrões de evolução temporal da história e a correlação dos acontecimentos para apoiar na decisão de publicar um livro.

REFERÊNCIAS

ALBERT, R.; BARABASI, A.-L. Statistical mechanics of complex networks. **Reviews of Modern Physics**, v. 74, p. 48–98, 2002.

AMANCIO, D. R. **Classificação de textos com redes complexas**. 2013. 302 f. Tese (Doutorado em Física Aplicada) — Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2013. Available at: http://www.teses.usp.br/teses/disponiveis/76/76132/tde-20012014-092439/. Access at: 6 jan. 2022.

AMANCIO, D. R. Authorship recognition via fluctuation analysis of network topology and word intermittency. **Journal of Statistical Mechanics: Theory and Experiment**, v. 2015(3):P03005, 2015.

AMANCIO, D. R. A complex network approach to stylometry. **PLoS ONE**, v. 10(8):e0136076, 2015.

AMANCIO, D. R. *et al.* Comparing intermittency and network measurements of words and their dependence on authorship. **New Journal of Physics**, v. 13:123024, 2011.

AMARAL, L. A. N.; OTTINO, J. M. Complex networks. **European Physical Journal B**, v. 38, p. 147–162, 2004.

ARRUDA, H. F. de; COSTA, L. da F.; AMANCIO, D. R. Using complex networks for text classification: Discriminating informative and imaginative documents. **EPL (Europhysics Letters)**, v. 113(2):28007, 2016.

ARRUDA, H. F. de *et al.* Paragraph-based representation of texts: A complex networks approach. **Information Processing & Management**, v. 56(3), p. 479–494, 2019.

ARRUDA, H. F. de *et al.* Representation of texts as complex networks: a mesoscopic approach. **Journal of Complex Networks**, v. 6(1), p. 125–144, 2018.

ASHOK, V. G.; FENG, S.; CHOI, Y. Success with style: using writing style to predict the success of novels. **Proceedings of the 2013 conference on empirical methods in natural language processing**, p. 1753–1764, 2013.

CARMI, E.; OESTREICHER-SINGER, G.; SUNDARARAJAN, A. Is oprah contagious? identifying demand spillovers in online networks. **NET Institute Working Paper**, v. 10-18, 2012.

CLEMENT, M.; PROPPE, D.; ROTT, A. Do critics make bestsellers? opinion leaders and the success of books. **Journal of Media Economics**, v. 20, p. 77–105, 2007.

CONG, J.; LIU, H. Approaching human language with complex networks. **Physics of Life Reviews**, v. 11(4), p. 598–618, 2014.

FELDMAN, R. Techniques and applications for sentiment analysis. **Communications of the ACM**, v. 56(4), p. 82–89, 2013.

FREEMAN, L. C. A set of measures of centrality based on betweenness. **Sociometry**, v. 40(1), p. 35–41, 1977.

FREEMAN, L. C. Centrality in networks: I. conceptual clarification. **Social Networks**, v. 1, p. 215–239, 1979.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. **arXiv:1405.4053**, 2014.

MANNING, C. D. *et al.* The stanford corenlp natural language processing toolkit. **Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, p. 55–60, 2014.

NAKAMURA, L. "words with friends": socially networked reading on goodreads. **PMLA**, v. 128(1), p. 238–243, 2013.

NEWMAN, M. E. J. The structure and function of complex networks. **SIAM**, v. 45(2), p. 167–256, 2003.

PROJECT Gutenberg. [*S.l.: s.n.*]. Available at: https://www.gutenberg.org/. Access at: 6 jun. 2022.

REVENUE of the U.S. book publishing industry 2008-2019. Statista. Available at: https://www.statista.com/statistics/271931/revenue-of-the-us-book-publishing-industry/. Access at: 6 out. 2021.

SHEHU, E. *et al.* The influence of book advertising on sales in the german fiction book market. **J Cult Econ**, v. 38, p. 109–130, 2013.

SOUZA, B. C. e *et al.* Accessibility and trajectory-based text characterization. **arXiv:2201.06665v1**, 2022.

WANG, X. *et al.* Success in books: predicting book sales before publication. **EPJ data science**, v. 8, p. 1–20, 2019.

WATTS, D. J. **Six Degrees**: The science of a connected age. New York: W.W. Norton & Company, 2003.