UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Classificação de Textos Curtos usando Redes Heterogêneas de Informação

Aline Naomi Arakaki

TCC do MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Aline Naomi Arakaki

Classificação de Textos Curtos usando Redes Heterogêneas de Informação

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. Ricardo Marcacini

USP – São Carlos
de 2022

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Arakaki, Aline Naomi
Ac Classificação de Textos Curtos usando Redes Heterogêneas de Informação / Aline Naomi Arakaki; orientador Ricardo Marcacini. -- São Carlos, 2022.
45 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. inteligência artificial. 2. redes heterogêneas. 3. aprendizado semisupervisionado. I. Marcacini, Ricardo , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Aline Naomi Arakaki

Classification of Short Texts using Heterogeneous Information Network

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 requitements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration Area : Artificial Intelligence

Advisor: Prof. Dr. Ricardo Marcacini

USP – São Carlos
2022

RESUMO

ARAKAKI, A. N. **Classificação de Textos Curtos usando Redes Heterogêneas de Informação**. 2022. 45 p. TCC (em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

Com o crescimento da produção de dados textuais, realizar a organização e gerenciamento desses dados de forma manual tornou-se inviável. Uma das soluções utilizadas ocorre através do aprendizado de máquina.

Em particular, o presente trabalho tem como foco avaliar a representação de redes heterogêneas de informação para textos curtos no cenário de weak labels. Essa metodologia é bastante utilizada por ser semisupervisionada e apresentar boa performance mesmo considerando o cenário de poucos dados rotulados sendo bastante relevante para aplicações reais.

Palavras-chave: Inteligência Artificial, redes heterogêneas, aprendizado semisupervisionado.

ABSTRACT

ARAKAKI, A. N. **Classification of Short Texts using Heterogeneous Information Network**. 2022. 45 p. TCC (em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

With the growing of textual data production, organizing and managing this data manually has become unfeasible. One of the solutions is machine learning. In particular, the present work focuses on evaluating the representation of heterogeneous information networks for short texts in the weak labels scenario.

This methodology is widely used because it is semi-supervised and performs well even considering little labeled data, which is quite relevant for real applications.

Keywords: Artificial intelligence, heterogeneous networks, semi-supervised learning.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Atualmente, vemos um aumento na produção de dados textuais. Esse crescimento pode ser encontrado em tweets, e-mails, relatórios, dentre outros. Entretanto, conseguir gerenciar, organizar esses dados de forma manual requer um grande esforço tornando essa atividade inviável. A análise, organização e gerenciamento torna as empresas mais competitivas no mercado e uma das formas encontradas para classificação massiva de textos e de rótulos ocorre através do aprendizado de máquina.

Dessa forma, com o crescimento de dados textuais vemos um crescente aumento de textos em formato curto, geralmente encontrado em análises de sentimentos, tweets e comentários. Em diversos cenários, como por exemplo análise de sentimentos, rotular esses dados para depois classificá-los pode levar muito tempo. Isso ocorre dado o volume e a necessidade de um recurso humano, podendo também necessitar de um expert sobre determinado assunto.

Entretanto, para realizar a classificação desse tipo de texto, verificamos dois principais desafios: a semântica esparsa dos textos e os ruídos ou incerteza nos dados anotados para treinamento.

No primeiro desafio, a semântica esparsa pode causar ambiguidade, além de consequentemente apresentar falta de contexto. Em alguns métodos é proposto incorporar informações adicionais, entretanto, apresentam o empecilho de manter os dados de forma relacional e com a mesma semântica.

O segundo desafio ocorre quando há ruídos ou incerteza nos dados anotados para treinamento de modelos, assim os modelos supervisionados tradicionais se tornam ineficientes (YANG T; HU, 2021). Esse desafio também é nomeado como weak labels. As weak labels são compostas por três formas. A incomplete supervision no qual grande parte dos dados não possuem classificação. O segundo tipo é o inexact supervision onde apenas coarse-grained são rotuladas, ou seja, os rótulos possuem classificações mais gerais e menos específicas. E por fim a inaccurate supervision que apresenta classificações não exatas, possuem verdades fundamentais

ou que podem ser difíceis de serem classificadas. (ZHOU, 2017) (SHI C; LI, 2017) (GONZÁLEZ J; INZA, 2015)

Uma forma promissora de lidar com os desafios acima é utilizar uma representação baseada em redes heterogêneas de informação. Nessa rede, documentos e seus atributos (e.g. termos e expressões) são considerados vértices e as arestas indicam relacionamentos do tipo documento-documento, documento-termo e termo-termo. Desse modo, a rede heterogênea ajuda a lidar com textos curtos, por considerar as relações como informação complementar. Ainda, alguns vértices da rede podem conter informação rotulada que é utilizada para classificar outros nós sem rótulos. Métodos de classificação em redes são interessantes nesse cenário, pois aprendem a importância de um vértice rotulado e lidar com weak labels.

Além disso, trabalhos recentes na literatura demonstram que redes heterogêneas de informação são representações mais complexas dos dados e permitem naturalmente adicionar novo conhecimento por meio de relações. (SHI C; LI, 2017) Métodos de aprendizado em redes heterogêneas, como LPHN (Label Propagation through Heterogeneous Network) e suas extensões, também definem níveis de importância para a informação rotulada.

Assim, esse projeto visa abordar a representação de redes heterogêneas de informação para textos curtos através da análise de sentimento. Além de avaliar um método de classificação de textos curtos representados por meio de redes heterogêneas de informação, será considerado o cenário de weak labels. Também será investigado estratégias para pré-processamento e representação de um conjunto de textos curtos por meio de redes heterogêneas. Além de métodos de aprendizado de máquina para classificação em redes heterogêneas de informação e a avaliação do impacto de diferentes níveis de weak labels nos métodos de classificação em redes heterogêneas.

1.1 Justificativa

Dado ao alto volume de textos produzidos atualmente, foi identificado a necessidade de aprofundar o tema para textos curtos, principalmente por apresentarem características específicas. Uma das dificuldades para classificação é a pequena quantidade de palavras que podem não apresentar uma boa mensuração de similaridade. Além de ocorrer um desbalanceamento entre a quantidade de dados e o próprio tamanho do texto.

Além disso, como textos curtos são enviados de forma rápida estes apresentam ampla quantidade de informações, assim há o empecilho de classificar esses textos manualmente. Geralmente, não apresentam padrão, podendo conter muitos erros ortográficos e ruídos. Os métodos tradicionais de aprendizado de máquina para textos curtos acabam não apresentando uma boa performance, com isso utilizaremos a metodologia de redes heterogêneas de informação.

Trabalhos recentes na literatura demonstram que redes heterogêneas de informação são representações mais complexas dos dados e permitem naturalmente adicionar novo conhecimento

por meio de relações (SHI C; LI, 2017). Métodos de aprendizado em redes heterogêneas, como LPHN e suas extensões, também definem níveis de importância para a informação rotulada. Um dos aspectos a ser investigado neste projeto é que os textos curtos também possuem falhas no processo de rotulação (e.g. weak labels), justificando a investigação de tais métodos (XIANG L; KAO B; ZHENG, 2016).

A exploração e avaliação dos modelos de representações de redes heterogêneas serão realizadas através da análise de sentimento. Neste caso, o modelo classificará se uma opinião é positiva ou negativa. Diferente de fatos, as opiniões e sentimentos são subjetivos. Dessa forma, é necessário examinar coleção de opiniões e de diferentes pessoas.

Geralmente, quando se trata de tweets, reviews, dentre outros por serem textos com poucas informações relevantes, se torna mais fácil classificar o sentimento enquando que para discussões sociais e politicas o tema se torna mais complexo podendo conter expressões de sarcarmos e ironia.

1.2 Objetivos

Esse projeto tem como finalidade desenvolver e avaliar modelos de representações de redes heterogêneas de informação para textos curtos. Assim, visa abordar o estudo de uma classificação semissupervisionada que possibilita o aprendizado a partir de uma combinação de dados não rotulados e rotulados. Dessa forma, tem-se como objetivo:

1. Desenvolver e avaliar dois modelos de redes heterogêneas: a primeira rede é uma rede bipartida que considera a relação entre documento e conceito e a segunda rede é uma extensão da citada anteriormente com a adição da relação entre conceito-conceito formando uma rede doc-conceito e conceito-conceito.

2. Avaliação das redes propostas anteriormente para um domínio de um conjunto de textos curtos, além da análise de impacto e comportamento de diferentes cenários de weak labels.

1.3 Organização do trabalho

Com relação à organização do trabalho o Capítulo 2 contém o referêncial teórico introduzindo conceitos gerais para textos como representações de texto, redes de informação e classificação de textos curtos. No Capítulo 3 detalha a apresentação dos modelos propostos das redes e às etapas do processo utilizado desde a explicação da base de dados, pré-processamento, modelagem das redes e critérios de avaliação. No Capítlo 4 demonstra a análise e avaliação experimental dos resultados para cada modelo. No Capítulo 5 expõem a conclusão e trabalhos futuros. E por fim, o Apêndice A que detalhadas as medidas de performance de avaliação do

segundo modelo apresentado considerando diferentes cenários de weak labels e limiares de similaridade de cosseno na relação entre conceito-conceito.

CAPÍTULO

2

FUNDAMENTOS E TRABALHOS RELACIONADOS

2.1 Representação de texto

Nesta sessão serão apresentados dois tipos de representações de textos: modelo espaço-vetorial e redes de informação.

2.1.1 Modelo Espaço-Vetorial

Os modelos de espaço-vetorial são frequentes no Aprendizado de Máquina, em que cada instância do conjunto de dados é representado por um vetor de característica (ZAKI MOHAMMED J; MEIRA, 2020).

Cada dimensão desse vetor equivale a uma característica ou atributo do conjunto de dados. Essas representações também podem ser utilizadas para classificação de textos (CHARU, 2018). Nesse caso, utiliza-se a nomenclatura "termo" para representar as dimensões que são formadas através das palavras do texto, podendo ser uma palavra simples, conjunto ou sequência de palavras.

O modelo espaço-vetorial pode ser representado através do conjunto de N documentos $D = d_1, d_2, ..., d_N$ e o conjunto de M termos que integram uma coleção de textos denominada $T = t_1, t_2, ..., t_M$. Dessa forma, os N vetores dos documentos de uma coleção são representados por M dimensões. Sendo assim forma-se uma matriz documento-termo (TAN P; STEINBACH, 2016). Essa matriz representa a união dos vetores das representações dos documentos de uma coleção.

Para modelos supervisionados há um atributo especial que indica a classe dos documentos. Em problemas de classificação, esse atributo é um campo de valor categórico, conforme ilustrado na Tabela 1.

Tabela 1 – Ilustração de uma matriz documento-termo.

-	t_1	t_2	t_3	...	t_{M-1}	t_M	Classe
d_1	$w_{d1,t1}$	$w_{d1,t2}$	$w_{d1,t3}$	...	$w_{d1,tM-1}$	$w_{d1,tM}$	c_{d1}
d_2	$w_{d2,t1}$	$w_{d2,t2}$	$w_{d2,t3}$	...	$w_{d2,tM-1}$	$w_{d2,tM}$	c_{d2}
d_3	$w_{d3,t1}$	$w_{d3,t2}$	$w_{d3,t3}$	...	$w_{d3,tM-1}$	$w_{d3,tM}$	c_{d3}
...	...	...	...		...	...	...
d_{N-1}	$w_{dN-1,t1}$	$w_{dN-1,t2}$	$w_{dN-1,t3}$	...	$w_{dN-1,tM-1}$	$w_{dN-1,tM}$	c_{dN-1}
d_N	$w_{dN,t1}$	$w_{dN,t2}$	$w_{dN,t3}$	...	$w_{dN,tM-1}$	$w_{dN,tM}$	c_{dN}

Na tabela acima verifica-se os pesos dos termos para definir a importância do termo para documentos. Essas medidas são quantitativas e são embasadas na frequência que termo t aparece no documento. Dessa forma, há três tipos principais de pesos encontrados para atribuir o valor $w_{di,tj}$. O primeiro é o caso binário onde os valores podem ser apenas 1 ou 0, sendo 1 quando o termo aparece no documento. Também ocorre o tf (do inglês term frequency - tf) que representa a frequência do termo, ou seja, quantas vezes o termo aparece no documento. E por fim, o tf-idf que realiza a frequência do termo ponderada pela inversa da frequência de documentos (do inglês term frequency - inverse document frequency - tf-idf) (CHARU, 2018).

Uma etapa relevante na geração da matriz documento-termo é o pré-processamento da coleção de documentos, visando a extração e geração de termos para obter uma representação estruturada dos textos.

Para a classificação de textos automática geralmente gera-se uma coleção de documentos com a matriz documento-termo, essa coleção nomeamos como bag-of-words. Vale destacar que Bag-of-Words ocorre quando utilizamos tokens simples como atributos e frequência como mecanismo de ponderação dos termos.

Esta representação tem como característica a alta dimensionalidade, dado a grande quantidade de palavras na coleção de texto e a alta esparsidade gerada por parte das palavras apresentarem baixa frequência nos documentos. Além disso, devido às suas principais características o algoritmo possui menor desempenho (YU L; LIU, 2004).

Outra característica do bag-of-words é não apresentar relações entre os termos, como por exemplo similaridade ou ordem. Caso seja necessário, é preciso representar as relações de forma explícita. Dessa forma, uma Bag-of-Words assume que os termos são independentes entre si (C. AGGARWAL; ZHAI, 2012).

Considerando a representação entre os termos, são utilizadas duas estratégias: frases estatísticas e sintáticas (SIDOROV, 2019). Para as frases estatísticas, também conhecidas como n-gramas, preserva-se a relação entre a ordem das palavras, como por exemplo as palavras “Estados” e “Unidos” que quando referido a país podem ser agrupadas para “Estados Unidos”. Enquanto que nas frases sintáticas utiliza-se qualquer conjunto de palavras que atenda alguma relação sintática ou que forme estruturas sintáticas específicas, como por exemplo a junção de

um adjetivo seguido de substantivo “impacto econômico”.

Mesmo utilizando tipos de frases, é necessário aplicar técnicas de pré-processamento pois estas auxiliam na manutenção e melhoria da qualidade do modelo. Além disso, pode ser utilizada para redução da dimensionalidade e/ou esparsidade das representações do modelo espaço-vetorial.

Há diversos tipos de técnicas de pré-processamento, alguns que se destacam são padronização das palavras, extração de palavras irrelevantes ou ruídos, agrupamento de palavras contendo o mesmo significado em apenas um atributo ou seção de palavras com uma mesma determinação e função sintática (CHARU, 2018).

Para a remoção de palavras podemos nomear como remoção de stopwords. Essas palavras são consideradas insignificantes para o algoritmo. Geralmente são pronomes, artigos, preposições e interjeições. Além disso, dependendo do contexto pode-se acrescentar mais palavras que sejam irrelevantes para o modelo.

Outra técnica utilizada é a simplificação de palavras que possuem o mesmo significado porém divergem pelo número, gênero ou tempo verbal. Há duas formas mais comuns de realizar a simplificação de palavras: radicalização e lematização. Na radicalização ocorre a redução das palavras ao seu radical, enquanto na lematização ocorre a união de diversos termos em um lema no qual altera-se verbos para o infinitivo e substantivos e adjetivos para o masculino singular. O objetivo é que essas palavras apresentem comportamento semântico e/ou sintático da palavra original. Na literatura frequentemente são utilizados mais a técnica de radicalização. Segundo Joachims (1999), os termos gerados pela radicalização das palavras formam representações mais compactas e sem que ocorra grande perda de informação.

Para redução do número de termos é possível utilizar o thesaurus, um dicionário no qual um termo é utilizado para substituir outros termos como por exemplo: “laranja”, "maçã"e “banana” são substituídos por “frutas” e dicionários de domínio em que termos são contidos em um dicionário específico.

Por fim, outra técnica utilizada é a relação entre termos no modelo espaço-vetorial. Pode ser criado um tipo de informação ou atributos auxiliares. Alguns exemplos são citações, hyperlinks. Entretanto, esse tipo de relação pode aumentar a dimensionalidade e a esparsidade das representações.

2.1.2 Redes de Informação

O conceito de redes é encontrado com diversas denominações podendo ser representada por “Uma rede é uma representação simplificada que reduz um sistema à uma representação abstrata” (NEWMAN, 2010) ou “Uma rede é um sistema de elementos que interagem ou regulam uns aos outros” ou um “conjunto de sistemas de coisas (objetos inanimados ou pessoas)” (BLANCO R;LIOMA, 2012).

As redes são compostas por “objetos” para representar os elementos e “relações” para caracterizar a ligação entre os elementos. Segundo alguns autores, os termos “grafos” e redes são utilizados para definir o mesmo conceito. Assim, são definidos como representações naturais entre grafos e objetos indicando relações formadas por um processo automático (MIHALCEA R; RADEV, 2011), enquanto há definições em que grafos funcionam como uma representação matemática das redes ou uma representação visual das iterações entre os componentes (NEWMAN, 2010).

As redes podem representar diversos sistemas como por exemplo: redes tecnológicas, redes sociais, redes sociais dentre outras. As redes tecnológicas indicam os objetos como equipamentos tecnológicos e as relações como a conexão estabelecida por esses objetos. As redes sociais podem corresponder às organizações, empresas, pessoas e as relações representadas pela amizade, comunicação e transação comercial. E por fim, as redes biológicas no qual os objetos são os elementos biológicos e as relações representa a interação entre esses elementos.

Ainda que existam diversos tipos de redes, elas podem ser representadas por $N =< O, R, W >$. Sendo O o conjunto de objetos da rede, R o conjunto das relações entre os objetos e W o conjunto de pesos das relações entre os objetos.

Observando os pares de objetos o_i, $o_j \in O$, as relações entre esses objetos é representada por r_{o_i,o_j}. Quando o sentido da relação é relevante mesmo não significando que há uma relação entre o_j e o_i denominamos como redes direcionadas. Enquanto que as redes não direcionadas não considera-se o sentido da relação.

Outra definição encontrada para classificação de grafos é feita através dos pesos de uma relação roi,oj. Quando os pesos são iguais, nomeamos como redes não ponderadas (do inglês *unweighted networks*) e no caso dos pesos serem diferentes denominamos redes ponderadas. As redes ponderadas são mais frequentes no aprendizado de máquina e em geral são aplicados valores positivos reais como os pesos das arestas.

Além disso, definimos se uma rede é homogênea ou heterogênea dependendo da quantidade de O por tipos de objetos, alguns exemplos são: redes de citações e redes de páginas web. Caso O possua um tipo de objeto classificamos como rede homogênea, “redes monodimensionais”, “redes unimodais” ou “redes mono-tipo”. Quando O é constituído por h diferentes tipos de objetos denominados redes heterogêneas, alguns exemplos são redes sociais e redes bibliográficas que contém livros, autores, locais de publicação. As redes heterogêneas também podem ser encontradas pelos termos “rede multitipo”,”redes multi-dimensionais”, “redes multimodais” ou “redes multirelacionais” para redes heterogêneas (SUN YIZHOU; HAN, 2012).

Um tipo especial de redes para representação de dados textuais são redes heterogêneas que são objetos de estudo deste projeto (YIZHOU S; HAN, 2013). As redes heterogêneas começaram a ganhar espaço na representação das coleções de textos. Um procedimento utilizado por Aery e Chakravarthy é a rede estrela no qual um objeto no centro caracteriza um documento

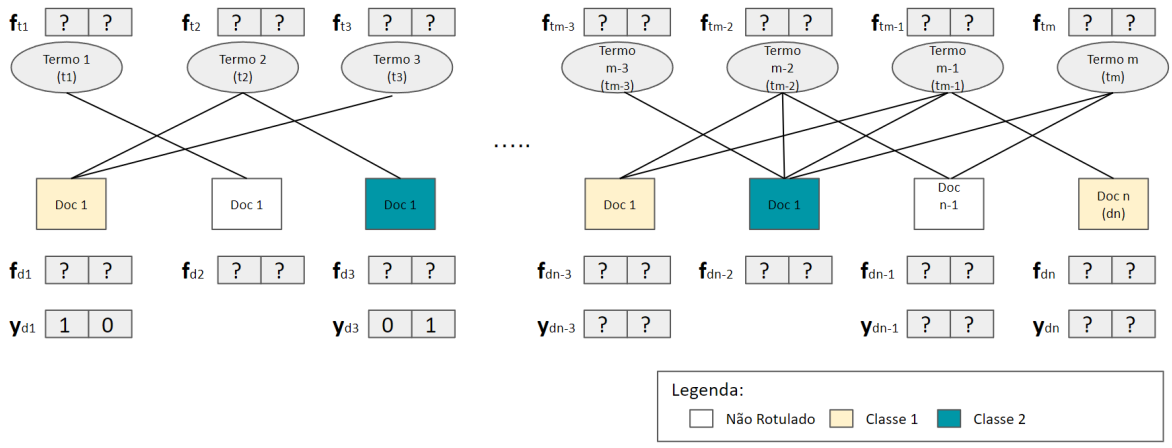
e as palavras desse documento se interligam a esse objeto central. Cada aresta é rotulada conforme a seção do documento sendo possível extrair sub-redes que são mais frequentes, dessa forma mesmo redes que possuam arestas ou vértices diferentes podem ser similares (AERY M; CHAKRAVARTHY, 2005).

Além disso, define-se um valor de importância para cada sub-rede dado pela representação de cada classe. Essas sub-redes podem ser comparadas com a rede de um novo documento e assim a classe é definida conforme a importância dessas sub-redes encontradas.

Segundo AGGARWAL S; LI (2011), para a rede proposta considera-se o conjunto de objetos da rede é composto por documentos e termos. Além disso, gera-se uma rede semi bipartida através das relações entre documentos oferecidas por hyperlinks nomeadas como informação estrutural. A relação entre termos e documentos dada pela ocorrência de um termo em um documento é denominada informação de conteúdo. Há diversas relações entre os objetos através dos documentos e termos. Dessa forma, podemos ter as relações documento-documento, documento-termo e termo-termo.

Considera-se $O = D \cup T$ no qual os objetos das redes correspondem aos termos T e documentos D. As redes heterogêneas e a relação entre documento-termo podem ser representadas conforme imagem abaixo:

Figura 1 – Ilustração para classificação de redes heterogênea bipartida.Adaptado de ROSSI (2015).



Considera-se f como um vetor de peso do objeto para cada classe determinada. Neste exemplo teremos os pesos para classe 1 e 2 obtida através do algoritmo após o processo de aprendizado. Por meio dos valores de f encontrados em cada termo (t) e documento (d) os vetores dos objetos são armazenados em uma matriz F. Na imagem são considerados a relação entre documento-termo. Também atribuímos a y_o peso dos rótulos que foram classificados previamente por um usuário ou especialista, ou seja, é a informação rotulada. Também foi criada uma matriz de informações Y contendo os dados reais. Note que uma matriz document-termo também pode ser interpretada como uma rede bipartida, na qual documentos e termos são vértices na rede e as

arestas indicam relacionamento entre eles.

2.1.3 Classificação de Textos Curtos

Conforme discutido anteriormente, há cada vez maior interesse na classificação de textos curtos, devido sua presença cada vez maior nas organizações na forma de e-mails, logs, queries, posts em redes sociais, entre outros (GE S; YE, 2014).

Textos curtos normalmente não apresentam uma sintaxe de escrita e geralmente oferecem um contexto limitado. Alguns exemplos são os tweets que não podem ultrapassar 140 caracteres e as queries que contém menos de 5 palavras.

Dessa forma, cresce o número de palavras que podem trazer ambiguidade. Para o entendimento humano é possível entender o contexto, entretanto isso acaba dificultando o aprendizado de máquina, especialmente tarefas de classificação de textos (CHEN L; XIU, 2022).

Assim, há um crescimento em novas abordagens para lidar com informações que possuem contexto limitado. Há dois tipos de abordagens: Explicit Representation Model (ERM) e Implicit Representation Model (IRM) (WANG Z; WANG, 2016).

No Explicit Representation Model (ERM) a análise e modelagem segue o processo tradicional da linguagem seguindo determinada taxonomia ou conhecimento base. Alguns exemplos de ERM são labeling (sense disambiguation) referente a identificação do significado correto de uma palavra ou sentença. A segmentation referente a divisão dos textos feita em palavras, sentenças, trechos, dentre outros. E syntax structure (dependency parsing) que busca identificar a relação entre termos. (HUA W; WANG, 2015).

A maior vantagem do ERM é resultar em representações que são simples de entender pelos humanos, além de ser customizado para diferentes cases. Do ponto de vista prático, o pré-processamento de textos é realizado de maneira clássica, conforme discutido nas seções anteriores. No entanto, para reduzir a ambiguidade, são incorporados recursos léxicos, como dicionários e taxonomias para selecionar termos mais relevantes ou incorporar termos sinônimos.

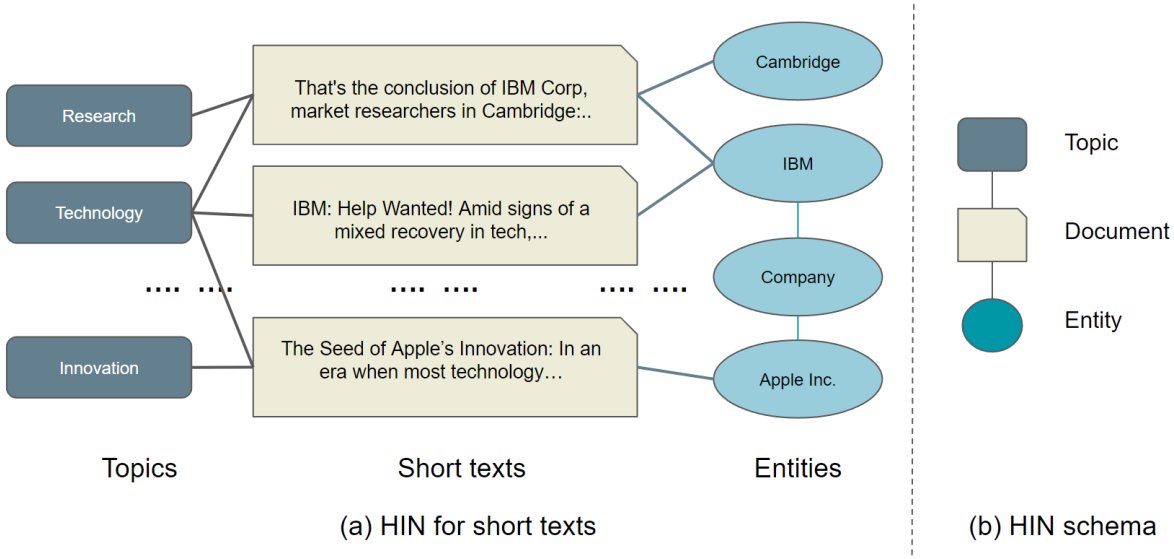
Já o Implicit Representation Model (IRM) busca aproveitar métodos recentes em deep learning para manter a semântica em textos curtos. Alguns exemplos são embedding por meio de word2vec e BERT (WANG; ZHOU; JIANG, 2020). O IRM apresenta a abordagem de utilizar o contexto baseado em uma palavra, frase ou sentença. São utilizados modelos como Redes Neurais Recorrente (RNN), Redes LSTM (Long Short Term Memory) e Transformers para incorporar representações pré-treinadas e realizar uma transferência de conhecimento visando lidar com textos curtos (LI W; LI, 2021).

Recentemente, uma abordagem híbrida visa unificar ERM e IRM por meio de redes heterogêneas de informação. A ideia é gerar relações explícitas enquanto que ao mesmo tempo incorpora word embeddings para o enriquecimento dos textos curtos.

Nos modelos apresentados utiliza-se a rede de informação (HIN) que busca integrar mais informações e capturar relações para melhorar a representação, bem como reduzir a ambiguidade Linmei *et al.* (2019). Apresentaram uma das primeiras abordagens nesse contexto, com foco em textos curtos.

Na Figura 2 é ilustrado um exemplo de rede heterogênea para textos curtos, em documentos de textos curtos, tópicos e entidades extraídas são vértices da rede heterogênea. Uma versão estendida deste trabalho, com experimentos mais robustos, foi publicado por Yang *et al.* (2021), confirmando os benefícios desta proposta.

Figura 2 – Exemplo de Rede Heterogênea para Classificação de Textos Curtos. Adaptado de Linmei *et al.* (2019)



Uma das limitações dessa abordagem é a necessidade de identificar tópicos e entidades dos dados textuais, que podem não estar presentes em textos formados por *queries* ou logs.

Dessa forma, nesse projeto busca-se uma estrutura mais simples para representação dos textos em redes heterogêneas, usando somente documentos e seus conceitos disponíveis.

Pode-se definir conceitos como uma sequência ordenada de termos que aparecem com recorrência em um determinado documento. O conceito também é determinado pela formação de N-gramas que tem como objetivo verificar palavras ou grupos de palavras comuns que aparecem em uma frequência estabelecida. Para cada conceito é esperado que preencha determinadas regras e uma uma frequência definida.

Assim, serão geradas duas redes: a rede documento-conceito e a rede documento-conceito e conceito-conceito.

A primeira rede é do tipo ERM enquanto a segunda há a possibilidade de gerar novas relações entre conceitos ou expandir os conceitos através de embeddings, por utilizar embeddings é considerado uma solução híbrida entre IRM e ERM.

CAPÍTULO

3

METODOLOGIA

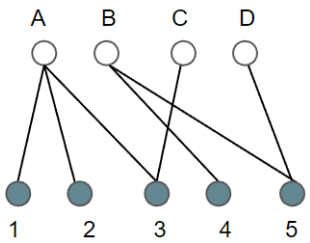
Nesta seção será apresentado o método proposto neste projeto, a descrição de cada etapa do processo, conforme descrito abaixo, além de trazer critérios de avaliação para classificação de textos curtos através de redes heterogêneas de informação considerando o cenário de weak labels. Pode-se considerar a metodologia dividida em três macro partes:

1. **Pré-processamento de dados:** nessa etapa foi explorado e analisado a base de dados além de compreender e identificar como os textos se comportam. Para tal estudo foi realizada a transformação desses dados para conceitos. E a utilização de técnicas de pré-processamento para textos.

2. **Modelagem de rede:** a modelagem será realizada por dois tipos de redes para a mesma base de dados:

   a) O primeiro modelo é uma rede bipartida que considera a relação documento e conceito. O conceito é formado através da etapa 3.1.

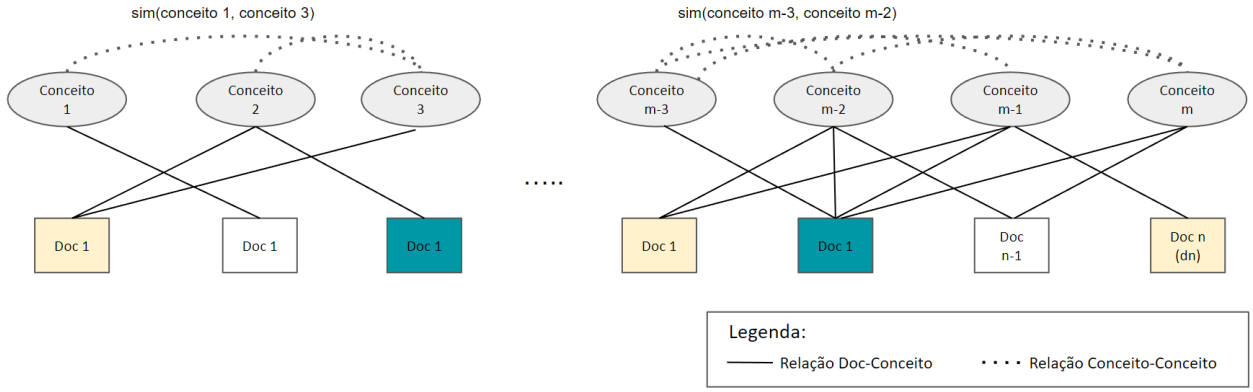
Figura 3 – Ilustração de rede bipartida



   b) O segundo modelo tem como objetivo criar uma rede baseada na primeira proposta adicionando a geração de links entre conceitos. O objetivo é que esses conceitos bem formados devem ser conectados com embedding similares. A embedding utilizada é

a pré-treinada do sentence-transformers que será explorada na subseção 3.1.4. E o critério para relação dos conceitos na rede será definido por um limiar de similaridade de cosseno.

Figura 4 – Ilustração de rede documento-conceito e conceito-conceito. Adaptado de ROSSI (2015).



3. **Caracterização da rede:** o objetivo é analisar e compreender as redes formadas. Foi definido o critério de avaliação, além de verificar os cenários de weak labels considerando uma tarefa de classificação semissupervisionada. Serão analisados resultados através da métrica f1-score macro para os modelos 1 e 2, citados anteriormente, alterando a quantidade de dados rotulados no treinamento.

3.1 Base de dados e preparação de dados

Nesta seção será apresentado a base de dados utilizada e como foi realizada a preparação de dados até a formação dos conceitos.

3.1.1 Base de dados

A base de dados utilizada neste presente trabalho é referente a ferramenta Google Photos (reviews). O Google Photos é um aplicativo gratuito que atua na nuvem e dentro dessa ferramenta é possível armazenar fotos, vídeos e capturas de telas do celular. É bastante utilizado devido a plataforma ser gratuita até determinado limite de armazenamento. É possível acessar as mídias em qualquer momento, convidar pessoas para compartilhar pastas de imagens e vídeos, dentre outras funcionalidades.

A base de dados escolhida retrata informações de comentários e as notas que foram atribuídas à plataforma considerando o intervalo de 1 a 5. Levando em consideração a base bruta, cerca de 72% obtiveram avaliação 5 e 7% avaliação 1, as demais permaneceram nas faixas entre 2 e 4. Dado que o intuito é verificar se os comentários possuem conotação positiva ou negativa decidiu-se seguir a avaliação de comentários que possuem apenas nota 1 e 5.

3.1.2 Descoberta de Conceitos Candidatos usando N-gramas

Esta etapa explora a formação de N-gramas e a descoberta de conceitos candidatos. Pode-se definir conceitos como uma sequência ordenada de termos que aparecem com recorrência em um determinado documento. O conceito também é determinado pela formação de N-gramas que tem como objetivo verificar palavras ou grupos de palavras comuns que aparecem em uma frequência estabelecida. Para cada conceito é esperado que preencha determinadas regras e uma uma frequência definida. As regras e frequência utilizadas neste trabalho serão apresentados na seção 3.1.3.

Referente à primeira etapa de formação de N-gramas, pode-se definir N-gramas como uma sequência contígua de N itens dentro de uma amostra textual. Dado o contexto, utilizaremos os comentários sobre a plataforma Google Photos e os N-gramas serão formados por palavras contidas nos comentários.

A geração dos N-gramas é feita através da criação de tokens e adotou-se tamanho de 4 a 11 tokens. Formando posteriormente uma base de dados que possui os candidatos e a frequência que esses aparecem nas frases.

Figura 5 – Ilustração de exemplo de N-gramas

Durante o processo da primeira etapa também foram utilizadas técnicas de preparação de dados, dentre elas, por exemplo: filtro para utilização de comentários com mais de 100 caracteres, verificação de *stopwords* no início e final de sentenças, dentre outras.

Através do método de pré-processamento de dados é possível verificar quais são os trechos de comentários que aparecem com maior frequência. Na base de dados utilizada o N-grama mais recorrente é “one of the best” comentando sobre o serviço oferecido. Dessa forma, através desse trecho percebe-se que são comentários que tendem a ser positivos sobre o Google Photos. Segue alguns exemplos:

"Google photos is very nice app to have and my experience has been a very good one !!!! Is one of the best to have that i no of thank you google photos !!!!!"

"Lovely it is incredible this is one of the best app ever it saves all my photos from day 1 to now iam totally speechless."

Além disso, também podem ser encontrados trechos que mostram percepções sobre o que é importante para a experiência do cliente. Dado os N-gramas mais frequentes foi encontrado questões relacionadas a “space on my phone”, “back up my photos”, “lost all my photos” e “get a new phone”. Dessa forma, é algo que pode ser levado em consideração na manutenção ou melhorias da ferramenta.

Entretanto, em alguns casos uns N-gramas não é possível identificar se a avaliação será positiva ou negativa conforme exemplos abaixo:

“I love how easy it is to use and i can backup all my photos and free up gigs and gigs of ***space on my phone*** *and easily access thr photos whenever i need.”*

“It is very frustrating that I have sent feedback and requested help MANY times, and yet no one has responded. Every time I go to free up ***space on my phone****, since getting this phone in February, it shuts the app down. I need to free* ***space on my phone****.”*

Dado a ambiguidade de significados é necessário também identificar, filtrar os N-gramas que são relevantes para compreender o sentimento dos clientes e realizar a seleção de conceitos.

3.1.3 Seleção de atributos

Após a etapa de formação de descoberta de conceitos candidatos, é necessário identificar quais conceitos são relevantes para a aplicação. Por se tratar de uma aplicação voltada à análise de sentimentos de reviews (HUTTO; GILBERT, 2014), decidiu-se utilizar a biblioteca *vaderSentiment - SentimentIntensityAnalyzer* que realiza a análise de sentimentos baseada em léxicos e regras de sentimentos expressos nas mídias sociais.

Através do *SentimentIntensityAnalyzer* é possível gerar um score de -1 a 1. Dessa forma, na documentação considera-se score com pontuação maior ou igual a 0.05 representam sentimentos positivos e menor que 0.05 e maior que -0.05, sentimentos neutro e menor que -0.05 sentimentos negativos.

Após gerado o score para cada N-grama seleciona-se apenas score superior a 0.1 ou inferior a -0.1 e que possuam frequência maior que 5. Sendo assim é possível extrair conceitos que são relevantes para avaliar e treinar os dados e retirar N-gramas que foram considerados neutros e com frequência menor que 5.

Para conceitos que possuem score acima de 0.1 pode-se encontrar exemplos como *"absolutely love this app"* e *"never have to worry"* enquanto para score menores encontra-se exemplos como *"lost all my photos"* e *"one of the worst"*.

3.1.4 Embeddings

Com o intuito de gerar a representação dos conceitos foi utilizada a técnica de embedding: **BERT Bidirectional Encoder Representations from Transformers**. O BERT foi treinado em

uma larga escala de textos que representam sentenças baseadas em um contexto. Este treinamento ocorreu em duas etapas: mascarar algumas palavras do texto e realizar a predição dela.

Na primeira etapa é dado uma sequência de texto e elaborado uma seleção randômica para substituir cerca de 15% das palavras por "máscaras". Consequentemente, o token das palavras selecionadas também são mascaradas. Esta etapa tem como objetivo fazer a rede neural prever a palavra mascarada. Assim, é possível realizar a exploração do contexto de palavras não mascaradas. A segunda etapa tem como intuito de prever qual será a próxima sentença e se essas sentenças são consecutivas e subsequentes.

A biblioteca utilizada no Python é a SentenceTranformers e a *'paraphrase-multilingual-mpnet-base-v2'*.

3.2 Redes Heterogêneas

Nesta seção serão apresentados dois tipos de redes heterogêneas, a primeira rede é uma rede bipartida que traz a relação entre documento e conceito. Dessa forma, será apresentado o funcionamento dessa rede e como ela é formada. O segundo modelo deriva do primeiro adicionando a relação conceito-conceito. Será explorado como foi realizada esta ligação através da utilização de embeddings e a medida de similaridade de cosseno.

No presente trabalho pode-se considerar que os documentos são representados pelos comentários do aplicativo Google Photos e os conceitos formados após a preparação de dados.

3.2.1 Redes bipartidas:documento-conceito

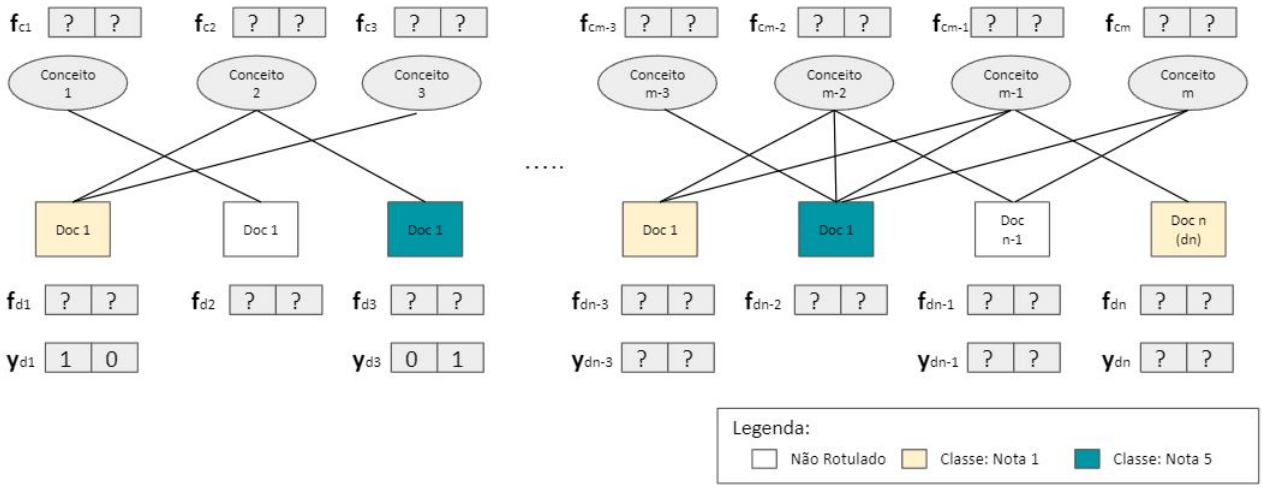
Inicialmente é proposto para classificação dos comentários uma rede bipartida da relação entre documento e conceito. Nesse caso, o documento é considerado o comentário dado à plataforma Google Photos e os conceitos são formados pela primeira etapa da metodologia.

O conjunto de rede é formado através das relações entre os documentos dada a ocorrência de um conceito em documento que é definido como informação de conteúdo. Através da geração desse modelo são atribuídos pesos do objeto para cada classe determinada. No presente trabalho utilizaremos pesos iguais, apenas considerando a relação entre arestas. Assim, atribuindo valores de f encontrados em cada documento e termo os vetores são armazenados em uma matriz. Para essa metodologia podemos apresentar a imagem 6.

Gaussian Fields and Harmonic Functions

O algoritmo *Gaussian Field and Harmonic Functions (GFHF)* é utilizado para classificação transdutiva em redes. A classificação transdutiva é gerada a partir de dados rotulados e não rotulados, assim é possível aprofundar as características do dado melhorando a performance do modelo para grafos.

Figura 6 – Ilustração de Rede Bipartida: documento-conceito



Esse algoritmo tem como base a utilização de campos gaussianos e funções harmônicas. A função harmônica é determinada através da média das classes dos objetos vizinhos ponderando os pesos das conexões. Sendo representada pela Equação 3.5:

$$\mathbf{f}_{o_i} = \frac{\sum_{o_j \in \mathscr{O}} w_{o_i,o_j} \mathbf{f}_{o_j}}{\sum_{o_j \in \mathscr{O}} w_{o_i,o_j}} \tag{3.1}$$

Os objetos que já possuem rótulo não sofrem alteração, a função harmônica ocorre apenas em objetos que não possuem rótulo. Assim, o algoritmo Gaussian Fields and Harmonic Functions pode ser considerado como um algoritmo de regularização que busca minimizar a função:

$$Q(F) = \frac{1}{2} \sum_{o_j \in \mathscr{O}} w_{o_i,o_j} (\mathbf{f}_{o_i} - \mathbf{y}_{o_i})^2 + \lim_{u \to \infty} u \sum_{o_j \in \mathscr{O}} (\mathbf{f}_{o_i} - \mathbf{y}_{o_i})^2 \tag{3.2}$$

Conforme descrito acima, os objetos rotulados não sofrem alteração e essa restrição é dada na equação através do $\lim_{u \to \infty}$.

Além disso, a equação 3.2 também pode ser utilizada para outros fins representados em outros tipos de rede.

Utiliza-se o algoritmo Label Propagation (LP) para minimizar iterativamente a função 3.2. Essa minimização ocorre através da equação matricial $F = PF$ iterativamente até a convergência dos valores da matriz F. Em relação a matriz P, considera-se tamanho $|O|x|O|$ em que uma célula p_{o_i,o_j} abrange a probabilidade de conexão entre objeto o_i e o_j caracterizada por:

$$P = D^{-1}W \tag{3.3}$$

Explorando o conceito de Class Mass Normalization (CMN), este pode ser representado pela divisão da informação de classe de um objeto o_i pela classe c_j, essa divisão utiliza todas

as informações dos objetos para mesma classe. E dessa forma, reduz as classes desbalanceadas. O CMN é usado para pós-processar as informações que foram geradas das classes através da solução iterativa. Aplicado o arg-max, segue a equação CVM 3.4 para classe dos objetos:

$$class(o_i) = arg \max_{\{c_l \in \mathscr{C}\}} Pr[c_l].\frac{f_{oi,cl}}{\sum_{o_j \in \mathscr{O}} f_{oj,cl}} \tag{3.4}$$

Retornando ao conceito de função harmônica, a mesma pode ser definida como uma caminhada aleatória na rede, considerando que o objeto o_i irá se movimentar aleatoriamente até uma partícula j. Essa caminhada ocorre até o encontro de um objeto rotulado que é denominado por absorbing random walk. Dessa forma, o algoritmo GFHF corresponde às informações de classe que é dada através da classe obtida em o_i e da caminhada aleatória.

Tendo em vista a explicação do algoritmo GFHF apresenta-se a extensão desse algoritmo utilizado para redes heterogêneas denominado LPHN. O *Label Propagation through Heterogeneous Network* realiza a classificação transdutiva sem que ocorra a necessidade de definição de parâmetros. Esse algoritmo é utilizado quando não é possível executar testes com muitos parâmetros ou que não ocorra entendimento desses parâmetros para o algoritmo.

3.2.2 Rede Documento-Conceito e Conceito-Conceito

Esta segunda abordagem é uma extensão da rede documento-conceito com a adição da relação entre conceitos. Assim, ambos possuem o mesmo método de propagação de rótulos descrito acima, o *Label Propagation through Heterogeneous Network* que atua como uma classificação transdutiva.

Para a criação dessa conexão entre conceitos, o primeiro passo foi transformar os conceitos em embeddings, descritos na seção 3.1.4. Ressaltando que na formação da rede bipartida não foi necessário utilizar esta técnica.

Em seguida, foi escolhida a medida de cosseno para verificar a similaridade entre eles. Esta medida é bastante utilizada em mineração de textos para indicar similaridade entre textos. A similaridade de cosseno é uma medida entre dois vetores em um determinado espaço vetorial trazendo o valor do cosseno do ângulo. Assim, quanto maior a métrica, maior a similaridade entre eles. Pode-se utilizar diversas medidas de similaridade para considerar se há conexão ou não entre esses conceitos. Assim, foi proposto a realização de diferentes *threshold* variando os limiares de similaridade de 0 a 1.

3.3 Critérios de Avaliação

Esse projeto visa abordar a representação de redes heterogêneas de informação para textos curtos. Além de avaliar um método de classificação de textos curtos representados por

meio de redes heterogêneas de informação, será considerado o cenário de *weak labels*.

Dessa forma, foram selecionadas aleatoriamente 30% da base de dados como teste. Dos 70% que seriam destinados para treino a fim de considerar cenários de weak labels, decidiu-se avaliar e comparar a performance do modelo variando a quantidade N de dados rotulados. Essa proporção geralmente costuma ser muito pequena em relação ao tamanho da base por se considerar uma classificação semi-supervisionada. Nesse caso utilizou-se dos 70% de dados de treino os seguintes cenários: 1%, 2%, 5%, 10%, 20% e 30% de dados rotulados.

Após a utilização do classificador, cada objeto na rede tem uma probabilidade para cada classe. Assim, considera-se a classe que possui maior probabilidade. A medida utilizada para verificar a performance do modelo em diferentes cenários é a f1-score macro.

Para explicação da métrica f1- score macro primeiramente são apresentados dois conceitos: Precisão-Macro e Revocação-Macro. A precisão nos dá a proporção de predições positivas que realmente eram positivas nomeadas como verdadeiros positivos (VP) em comparação a todas as predições positivas que são denominados verdadeiros positivos (VP) e falsos positivos (FP). Na Precisão- Macro adiciona-se C que representa o conjunto de classes do problema. Dessa forma, a métrica será a média aritmética da precisão para cada conjunto.

$$Pmacro = \frac{1}{|C|} \sum_{ci \in C} \frac{VPc_i}{VPc_i + FPc_i} \tag{3.5}$$

A segunda métrica, o revocação nos trás a proporção entre as predições positivas (VP) comparada a todos reais positivos, sendo elas denominadas verdadeiros positivos (VP) e falsos negativos (FN). Da mesma forma que ocorre com a precisão-macro, a revocação-macro considera o conjunto de classes do problema. Dessa forma, segue a equação 3.6 abaixo:

$$Rmacro = \frac{1}{|C|} \sum_{ci \in C} \frac{VPc_i}{VPc_i + FNc_i} \tag{3.6}$$

Assim, a métrica f1-score macro é a média harmônica entre a precision-macro e a revocação macro. Esta métrica é utilizada principalmente para penalizar as classes que são majoritárias, caso algumas das métricas sejam baixas entre precision e recall o f1-score ficará mais próxima do menor número. Além de ser utilizada para dados desbalanceados.

$$F1macro = \frac{2 * Pmacro * Rmacro}{Pmacro + Rmacro} \tag{3.7}$$

CAPÍTULO

4

RESULTADOS

Neste capítulo são apresentados os resultados dos experimentos do modelo 1 e modelo 2. No primeiro experimento serão explorados a medida do f1-score macro para diferentes cenários de nós rotulados e a análise de impacto ao variar a quantidade de nós rotulados para uma rede bipartida em que há uma relação documento e conceito.

No segundo experimento foi levado em consideração que além da formação dos conceitos também serão explorados a relação entre documento-conceito (modelo 1) e adicionalmente a relação de conceito-conceito. Na relação conceito-conceito foi utilizada a medida de similaridade cosseno para a definir se há conexão ou não entre os conceito e será apresentada a comparação entre os resultados obtidos da f1-score macro para diferentes cenários de quantidade de nós rotulados e medidas de similaridade.

4.1 Experimento 1

O primeiro experimento realizado utilizando rede bipartida da relação entre documento e conceito foi aplicado para classificar comentários, através de dados rotulados com as notas 1 e 5. Sendo 1 referente a comentários negativos e 5 aos positivos. Dessa forma, é possível prever através da base de comentários a possível avaliação que seria dada à plataforma Google Photos.

Na etapa de formação da rede bipartida encontraram-se os seguintes valores: 10563 vértices e 12266 arestas. Através da formação dessa rede foram realizados experimentos para avaliar o desempenho do modelo para cenário *weak labels*. Atualmente com a grande quantidade de informações dificulta-se o trabalho de rotulação de dados, pois, esta depende da classificação dada por um usuário ou especialista. Vendo essa necessidade foi criado cenários para verificar o impacto na quantidade de rótulos no modelo.

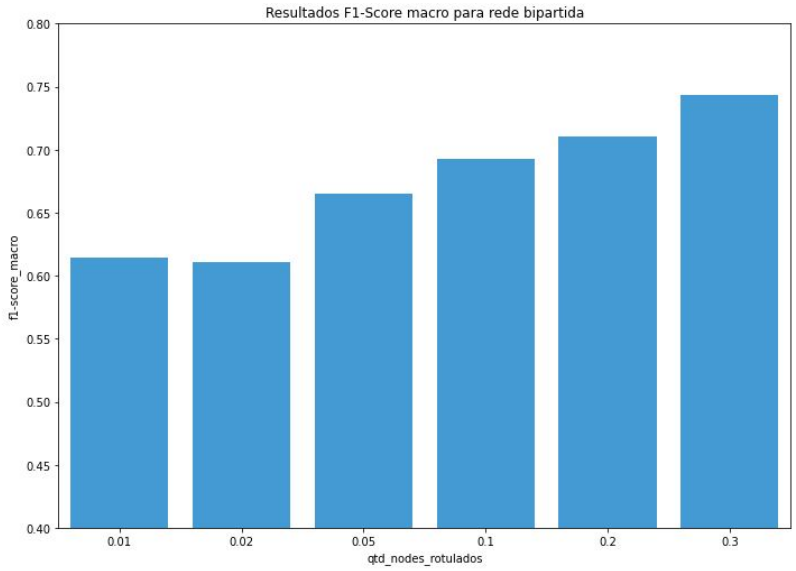
Na tabela 2 são apresentados os resultados do modelo de rede bipartida. Pode-se observar que o modelo proposto traz resultados satisfatórios mesmo com quantidades pequenas de dados

Tabela 2 – Resultados Modelo de Rede Bipartida.

-	$Cenario1$	$Cenario2$	$Cenario3$	$Cenario4$	$Cenario5$	$Cenario6$
$\#nodes\ rotulados$	(0.01)	(0.02)	(0.05)	(0.1)	(0.2)	(0.3)
$F1-Score_{macro}$	0.61	0.61	0.66	0.69	0.71	0.74

rotulados relatando que este método poderia ser utilizado para *weak label*. É esperado, conforme mostra a tabela, que quanto maior o número de labels há uma melhora na métrica f1-score.

Figura 7 – Gráfico de Performance para o Experimento 1



4.2 Experimento 2

No experimento 2 ao adicionar a relação entre conceitos realiza-se a comparação de performance para diversas medidas de similaridade de cosseno, assim será analisado qual o comportamento ao alterar esses intervalos e com isso, verificar o impacto do resultado da métrica proposta f1-score.

Dessa forma, foi gerado visões para analisar os resultados do modelo conforme demonstrado abaixo. Para entendimento, cada gráfico representa o resultado para determinado número de labels rotulados, assim como no experimento 1 utilizou-se dos 70% de dados de treino os cenários: 1%, 2%, 5%, 10%, 20% e 30% de dados rotulados.

Cada gráfico possui dois eixos y: quantidade de arestas e a métrica f1-score macro. Enquanto o eixo x é formado pela medida de similaridade de cosseno. Considerando os gráficos abaixo, foi definido que a medida de similaridade representa o limiar para realização de conexões. Por exemplo, quando é apresentada a similaridade de cosseno igual ao valor 0, isso significa que para realizar uma conexão entre nodes a similaridade deve ser igual ou maior que zero, nesse caso todos os outros conceitos obtiveram conexão. Quando a similaridade possui limiar de 0.3,

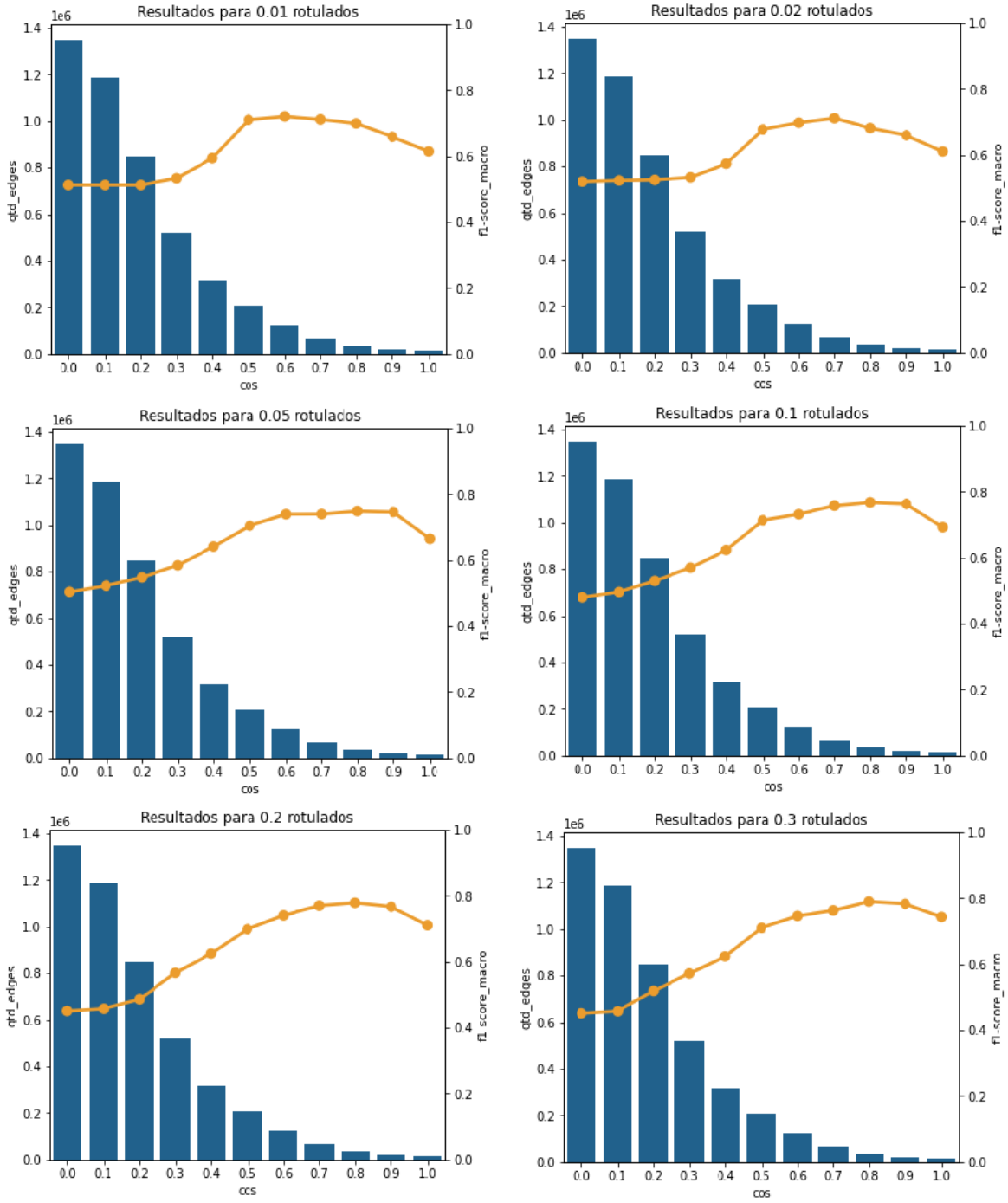


Figura 8 – Resultados Experimento 2: cada gráfico representa o resultado do modelo para diversos cenários de dados rotulados. As barras azuis representam a quantidade de arestas para cada medida de similaridade de cosseno no eixo x e a linha em laranja a métrica de desempenho f1-score macro.

por exemplo, significa que foi estabelecido que apenas ocorrerá conexão entre os conceitos se a similaridade por cosseno for maior que 0.3.

Também é possível verificar pelas barras azuis que quanto maior a similaridade estabelecida para realizar essa ligação, menor o número de arestas. Dessa forma, em todos os gráficos

ocorrem um decrescimento, pois, o limiar para realizar a conexão de arestas se torna mais restrito.

Observa-se que a performance da métrica f1-score começa a aumentar gradativamente, entretanto, em determinado momento restringir a criação de novas arestas ou conexão faz com que o modelo diminua sua performance.

Dessa forma, considerando casos que se aproximam da realidade foi estabelecido um limiar de cos = 0.5 ,pois, ele atua bem em 1% de dados rotulados. Neste resultado teremos 10712 nodes e 205275 arestas.

Assim, é possível verificar os resultados gerais para o limiar determinado. Conforme tabela 3:

Tabela 3 – F1-Score macro considerando o limiar de 0.5 para similaridade de cosseno

-	*Cenario*1	*Cenario*2	*Cenario*3	*Cenario*4	*Cenario*5	*Cenario*6
#nodes rotulados	(0.01)	(0.02)	(0.05)	(0.1)	(0.2)	(0.3)
$F1 - Score_{macro}$	0.71	0.68	0.70	0.71	0.70	0.71

Nesta escolha de similaridade percebe-se que mesmo aumentando o número de nodes rotulados a performance do modelo não possui variações bruscas e traz bons resultados com poucos dados rotulados.

Levando em considerações aplicações reais, normalmente, há poucos dados rotulados. Assim, para avaliar o desempenho e comparar os modelos foram levados principalmente em consideração o cenário 1 onde (0.01) dados são rotulados. O modelo 1 apresentou 0.61 para a métrica f1-score macro e 0.71 para o modelo 2.

Ao criar mais arestas através da relação conceito-conceito percebe-se que o segundo modelo tem melhor ganho de desempenho e competitividade. Para que o primeiro modelo tivesse desempenho semelhante ao segundo seria necessário ter cerca de (0.20) de dados rotulados. Assim, caso o objetivo fosse gerar novos dados rotulados haveria a necessidade de alocar recursos de tempo e especialista para possuir resultado similar ao modelo 2.

Os resultados experimentais completos, envolvendo todos os limiares estão disponíveis na Tabela A.

CAPÍTULO

5

CONCLUSÃO

Com o crescimento da produção textual, realizar as análises e gerenciamento desses dados de forma manual se torna uma atividade inviável. Dessa forma, uma das abordagens para lidar com a classificação de textos ocorre através do aprendizado de máquina.

Há diversos métodos utilizando este tipo de aprendizado e o presente trabalho abordou a classificação de textos curtos através de redes heterogêneas. Os modelos compostos por redes heterogêneas tem como vantagem serem semisupervisionado, ou seja, ele utiliza para treino tanto dados rotulados e não rotulados.

Isso foi refletido no primeiro experimento de rede bipartida, pois verificou-se um desempenho satisfatório mesmo com quantidades pequenas de dados rotulados relatando que este método tende a ser favorável para cenários de *weak label*.

No segundo experimento foi realizada a extensão do experimento 1. Além da relação estabelecida entre documento-conceito foi adicionada a relação entre os próprios conceitos através da similaridade de cosseno. Dessa forma, verificou-se que esta adição de conexão e a criação de mais arestas dentro do grafo melhorou o desempenho do modelo. Assim é possível destacar que esta abordagem utilizada possui bom desempenho mesmo com quantidades pequenas de dados rotulados. Também foi definido um limiar para realização das ligações entre as arestas dos conceitos.

Como limitação vale relembrar que os experimentos foram realizados para um tipo específico de base de dados, neste caso a avaliação de um aplicativo. Assim, as próximas etapas deste estudos incluem avaliação dos resultados para outros tipo de base de dados e a realização desse experimento para outros domínios como por exemplo classificação de logs sistemicas, comentários de NPS, dentre outras.

Esta abordagem também pode ser utilizada através da combinação entre treino e a supervisão de humanos. A rede gera um resultado e um humano confere as classificações e tendo

esses novos dados rotulados adicionamos e retreinamos esta rede.

REFERÊNCIAS

AERY M; CHAKRAVARTHY, S. Infosift: Adapting graph mining techniques for text classification. In: ____. **In Proceedings of the Florida Artificial Intelligence Research Society Conference**. [S.l.]: AAAI Press, 2005. p. 6, 7, 39, 46, 49, 50, 89. Citado na página 23.

AGGARWAL S; LI, N. On node classification in dynamic content-based networks. **Proceedings of the SIAM International Conference on Data Mining**, p. 5, 6, 49, 50, e 60, 2011. Citado na página 23.

BLANCO R;LIOMA, C. Graph-based term weighting for information retrieval. In: ____. **Graph-based term weighting for information retrieval**. [S.l.]: Information Retrieval, 2012. Citado na página 21.

C. AGGARWAL; ZHAI, C. C. Mining text data. In: ____. **Mining text data**. Boston/Dordrecht/London: Springer Science Business Media, 2012. Citado na página 20.

CHARU, A. Machine learning for text. In: ____. **Machine learning for text**. Nova York: Springer, 2018. Citado nas páginas 19, 20 e 21.

CHEN L; XIU, B. Z. Multiple weak supervision for short text classification. **Applied Intelligence**, p. 1–16, 2022. Citado na página 24.

GE S; YE, Y. H. X. B. S. Short text classification: A survey. **Journal of multimedia 9**, v. 5, p. 365, 2014. Citado na página 24.

GONZÁLEZ J; INZA, I. L. J. Weak supervision and other non-standard classification problems: A taxonomy. **ELSEVIER**, v. 69, n. 1, p. 50, 2015. Citado na página 16.

HUA W; WANG, Z. Short text understanding through lexical-semantic analysis. **International Conference on Data Engineering (ICDE)**, 2015. Citado na página 24.

HUTTO, C. V.; GILBERT, E. **A parsimonious rule-based model for sentiment analysis of social media text**. [S.l.]: In Proceedings of the international AAAI conference on web and social media, 2014. v. 8. 216-225 p. 1. Citado na página 30.

LI W; LI, L. Combining knowledge with attention neural networks for short text classification. In: ____. **Combining Knowledge with Attention Neural Networks for Short Text Classification**. [S.l.]: Springer, 2021. p. 240–251. Citado na página 24.

LINMEI, H.; YANG, T.; SHI, C.; JI, H.; LI, X. Heterogeneous graph attention networks for semi-supervised short text classification. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. [S.l.: s.n.], 2019. p. 4821–4830. Citado nas páginas 9 e 25.

MIHALCEA R; RADEV, D. Graph-based natural language processing and information retrieval. In: ____. **Graph-based natural language processing and information retrieval**. [S.l.]: Cambridge University Press, 2011. Citado na página 22.

NEWMAN, M. Networks: An introduction. In: ____. **Networks: An Introduction**. [S.l.]: Oxford University Press, 2010. Citado nas páginas 21 e 22.

ROSSI, R. Classificação automática de textos por meio de aprendizado de máquina baseado em redes. **Instituto de Ciências Matemática Computacional, Universidade de São Paulo**, p. 59–60, 2015. Citado nas páginas 9, 23 e 28.

SHI C; LI, Y. Z. J. S. Y. Y. P. A survey of heterogeneous information network analysis. **IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING**, v. 29, n. 1, p. 24–25, 2017. Citado nas páginas 16 e 17.

SIDOROV, G. Syntactic n-grams in computational linguistics. In: ____. **Syntactic n-grams in computational linguistics**. [S.l.]: Springer, 2019. Citado na página 20.

SUN YIZHOU; HAN, J. Mining heterogeneous information networks: principles and methodologies. In: ____. **Synthesis Lectures on Data Mining and Knowledge Discovery**. [S.l.]: Morgan Claypool, 2012. v. 2, p. 1–159. Citado na página 22.

TAN P; STEINBACH, M. K. V. Introduction to data mining. In: ____. **Introduction to Data Mining**. [S.l.: s.n.], 2016. v. 2. Citado na página 19.

WANG, S.; ZHOU, W.; JIANG, C. A survey of word embeddings based on deep learning. **Computing**, Springer, v. 102, n. 3, p. 717–740, 2020. Citado na página 24.

WANG Z; WANG, H. Understanding short texts. **Annual Meeting of the Association for Computational Linguistics**, p. 1–4, 2016. Citado na página 24.

XIANG L; KAO B; ZHENG, Y. H. Z. On transductive classification in heterogeneous information networks. **In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management**, p. 811–820, 2016. Citado na página 17.

YANG, T.; HU, L.; SHI, C.; JI, H.; LI, X.; NIE, L. Hgat: Heterogeneous graph attention networks for semi-supervised short text classification. **ACM Transactions on Information Systems (TOIS)**, ACM New York, NY, v. 39, n. 3, p. 1–29, 2021. Citado na página 25.

YANG T; HU, L. S. C. J. H. L. X. N. L. H. **Heterogeneous Graph Attention Networks for Semi-supervised Short Text Classification**. [S.l.], 2021. 1-5 p. Citado na página 15.

YIZHOU S; HAN, J. Mining heterogeneous information networks: a structural analysis approach. **Acm Sigkdd Explorations Newsletter 14**, v. 2, p. 20–28, 2013. Citado na página 22.

YU L; LIU, H. Efficient feature selection via analysis of relevance and redundancy. **Journal of Machine Learning Research**, v. 5, p. 19, 2004. Citado na página 20.

ZAKI MOHAMMED J; MEIRA, W. J. Data mining and machine learning: Fundamental concepts and algorithms. In: ____. **Data Mining and Machine Learning: Fundamental Concepts and Algorithms**. Cambridge: Cambridge University Press, 2020. cap. 3, p. 102–187. Citado na página 19.

ZHOU, Z. A brief introduction to weakly supervised learning. **National Science Review**, v. 5, n. 1, p. 44–53, 2017. Citado na página 16.

APÊNDICE

A

RESULTADOS EXPERIMENTAIS 2

Tabela 4 – Resultados Experimentais para Rede Doc-Conceito e Conceito-Conceito

#nodes rotulados	qtd nodes	cos	f1 score macro	qtd nodes
0.01	0	0.51	10736	1348400
0.01	0.1	0.51	10736	1184985
0.01	0.2	0.51	10736	850142
0.01	0.3	0.53	10736	521740
0.01	0.4	0.59	10735	317217
0.01	0.5	0.71	10712	205275
0.01	0.6	0.72	10704	125294
0.01	0.7	0.71	10693	67491
0.01	0.8	0.7	10685	37857
0.01	0.9	0.66	10664	18384
0.01	1	0.61	10563	12266
0.02	0	0.52	10736	1348400
0.02	0.1	0.52	10736	1184985
0.02	0.2	0.52	10736	850142
0.02	0.3	0.53	10736	521740
0.02	0.4	0.57	10735	317217
0.02	0.5	0.68	10712	205275
0.02	0.6	0.7	10704	125294
0.02	0.7	0.71	10693	67491
0.02	0.8	0.68	10685	37857
0.02	0.9	0.66	10664	18384
0.02	1	0.61	10563	12266
0.05	0	0.5	10736	1348400

Continuação na próxima página

Tabela 4 – continuação da página anterior

#nodes rotulados	qtd nodes	cos	f1 score macro	qtd nodes
0.05	0.1	0.52	10736	1184985
0.05	0.2	0.55	10736	850142
0.05	0.3	0.58	10736	521740
0.05	0.4	0.64	10735	317217
0.05	0.5	0.7	10712	205275
0.05	0.6	0.74	10704	125294
0.05	0.7	0.74	10693	67491
0.05	0.8	0.75	10685	37857
0.05	0.9	0.75	10664	18384
0.05	1	0.66	10563	12266
0.1	0	0.48	10736	1348400
0.1	0.1	0.5	10736	1184985
0.1	0.2	0.53	10736	850142
0.1	0.3	0.57	10736	521740
0.1	0.4	0.62	10735	317217
0.1	0.5	0.71	10712	205275
0.1	0.6	0.73	10704	125294
0.1	0.7	0.76	10693	67491
0.1	0.8	0.77	10685	37857
0.1	0.9	0.76	10664	18384
0.1	1	0.69	10563	12266
0.2	0	0.45	10736	1348400
0.2	0.1	0.46	10736	1184985
0.2	0.2	0.49	10736	850142
0.2	0.3	0.57	10736	521740
0.2	0.4	0.62	10735	317217
0.2	0.5	0.7	10712	205275
0.2	0.6	0.74	10704	125294
0.2	0.7	0.77	10693	67491
0.2	0.8	0.78	10685	37857
0.2	0.9	0.77	10664	18384
0.2	1	0.71	10563	12266
0.3	0	0.45	10736	1348400
0.3	0.1	0.46	10736	1184985
0.3	0.2	0.52	10736	850142
0.3	0.3	0.57	10736	521740

Continuação na próxima página

Tabela 4 – continuação da página anterior

#nodes rotulados	qtd nodes	cos	f1 score macro	qtd nodes
0.3	0.4	0.62	10735	317217
0.3	0.5	0.71	10712	205275
0.3	0.6	0.75	10704	125294
0.3	0.7	0.76	10693	67491
0.3	0.8	0.79	10685	37857
0.3	0.9	0.78	10664	18384
0.3	1	0.74	10563	12266


ICMC
USP
SÃO CARLOS