UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO

Karen da Silva Lopes

Análise Exploratória de tópicos em iniciativas ESG adotadas em empresas brasileiras através de Processamento de Linguagem Natural

São Carlos

2022

Karen da Silva Lopes

Análise Exploratória de tópicos em iniciativas ESG adotadas em empresas brasileiras através de Processamento de Linguagem Natural

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Afonso Paiva

São Carlos

2022

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

S856m

Lopes, Karen da Silva
Análise Exploratória de tópicos em iniciativas ESG adotadas em empresas brasileiras através de Processamento de Linguagem Natural / Karen da Silva Lopes ; orientador Prof. Dr. Afonso Paiva. – São Carlos, 2022.
50 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Ciências de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. Ciência de Dados. 2. Processamento de Linguagem Natural. 3. Tópicos. 4. LDA. I. Paiva, Afonso, orient. II. Título.

Karen da Silva Lopes

Análise Exploratória de tópicos em iniciativas ESG adotadas em empresas brasileiras através de Processamento de Linguagem Natural

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Data de defesa: 05 de março de 2022

Comissão Julgadora:

Prof. Dr. Afonso Paiva
Orientador

Professor
Convidado1

São Carlos
2022

AGRADECIMENTOS

A Deus, pois sem Ele eu não seria nada.

A minha mãe, por todo amor e incentivo aos meus estudos; sem o seu esforço eu não teria chegado até aqui. Obrigada também por seu meu exemplo. E ao meu pai por toda ajuda e suporte.

Ao meu noivo, Gilmar Neves, meu companheiro em todos os momentos que sempre acreditou em mim, até mesmo quando duvidei da minha capacidade. Obrigada por toda alegria, apoio, dedicação e compreensão durante essa etapa e também ao amigo PGil.

Ao meu orientador, Afonso Paiva Neto, por toda confiança no meu trabalho, pelos conselhos e paciência e contribuições para realização desse trabalho.

Agradeço a Fundação de Apoio à Física e à Química (FAFQ) pela concessão da bolsa.

A todos os professores e monitores com quem aprendi muito e fizeram um excelente trabalho durante todo o curso.

Finalmente, a todos que contribuíram de alguma forma para este trabalho.

Obrigada!

RESUMO

LOPES, K.S. **Análise Exploratória de tópicos em iniciativas ESG adotadas em empresas brasileiras através de Processamento de Linguagem Natural**. 2022. 50p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

A relevância de questões ambientais, sociais e de governança (ESG) tem sido cada vez mais incorporadas em instituições financeiras, empresas, governos e bolsas de valores. Muitas informações importantes são encontradas em relatórios, notícias e outros meios de publicidade que demonstram o desempenho de uma organização. Portanto, o uso adequado dessas informações pode trazer uma vantagem competitiva para o negócio. Neste cenário, o objetivo principal deste trabalho consiste no desenvolvimento de um modelo para extração de tópicos para as principais iniciativas ESG utilizando técnicas de mineração de texto, visualização de informação e Processamento Natural de Linguagem (PLN). Aplicou-se técnicas de mineração de texto como nuvem de palavras e bigramas foram utilizadas para realizar a análise exploratória dos dados. Além disso, o método de Latent Dirichlet allocation (LDA) foi utilizado e gerou o melhor modelo com 10 tópicos.

Palavras-chave: ESG, PLN, Modelagem de Tópicos, LDA.

ABSTRACT

LOPES, K.S. **Exploratory analysis of topics in ESG initiatives adopted in Brazilian companies through Natural Language Processing**. 2022. 50p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

The relevance of environmental, social and governance (ESG) issues has been increasingly embedded in financial institutions, corporations, governments and stock exchanges. The majority information is found in reports, news and other advertising media that demonstrate an organization's performance. Therefore, the proper use of this information can bring a competitive advantage to the business. In this scenario, the main objective of this work is to develop a topic classification model for the main ESG initiatives using text mining, information visualization and Natural Language Processing (NLP) techniques. Text mining techniques such as word clouds and bigrams were used to perform exploratory data analysis. In addition, the Latent Dirichlet allocation (LDA) method was used and generated the best model with 10 topics.

Keywords: ESG, NLP, Topic Modeling, LDA.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A economia mundial busca cada mais por investimentos sustentáveis e uma iniciativa crescente é a adoção de melhores práticas ESG (sigla em inglês para *Environmental, Social, and Corporate Governance*), que significa a adoção de questões de aspectos ambientais, sociais e de governança corporativa nas empresas (MISHRA, 2020; UNGARETTI, 2020). Estes pilares consolidados e difundidos como ESG surgiu em 2004 na publicação do Pacto Global [1] resultante da iniciativa liderada pela Organização das Nações Unidas (ONU). Neste relatório 20 instituições financeiras de 9 países se reuniram e desenvolveram diretrizes e recomendações sobre os fatores ESG e concluíram que a incorporação dessas práticas geravam mercados mais sustentáveis e melhores resultados para a sociedade.

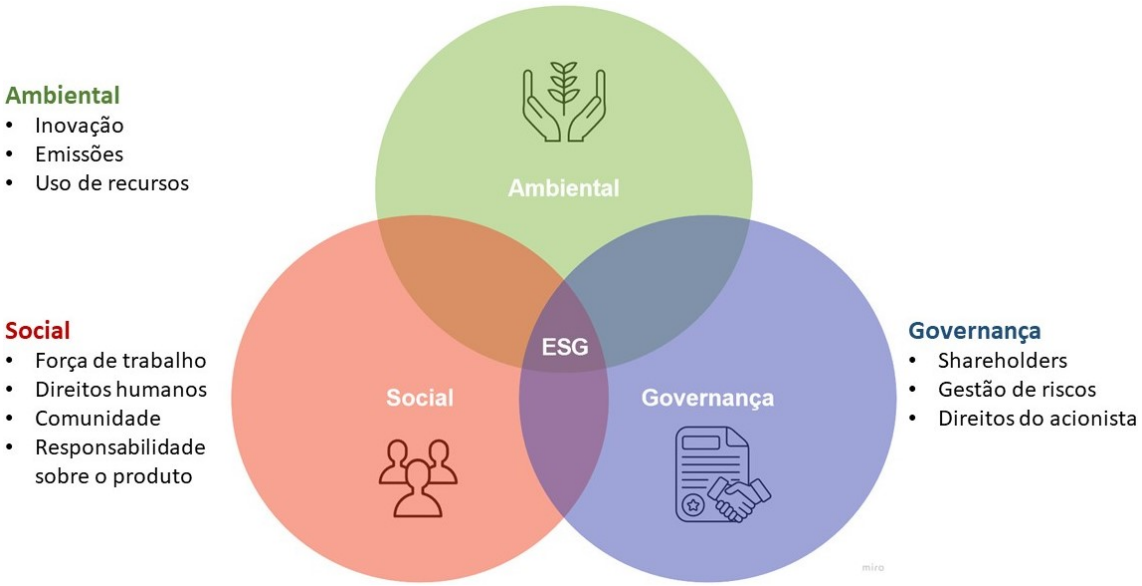
A sigla ESG no mundo de investimento busca olhar as empresas de forma holística com base nos critérios que incorporam questões ambientais, sociais e de governança. Sob a perspectiva ambiental é possível citar ações como redução de emissão de gases do efeito estufa, uso de recursos naturais, eficiência energética como medidas comuns praticadas pelas empresas. No aspecto social, iniciativas que promovem inclusão e diversidade, direitos humanos, relações com comunidades são recorrentes. Assim como, sob o pilar de governança é fundamental a ética e transparência das empresas.

Segundo Ungaretti (2020), essa tendência pode ser observada no mercado global, onde mais de US$30 trilhões em *ativos sob gestão* (AuM) estão em fundos que estabeleceram estratégias sustentáveis. Além disso, o número de signatários do código de sustentabilidade do PRI (*Principles for Responsible Investment*) ultrapassou a marca de 3.000 membros e totalizando de US$ 100 trilhões em AuM, o que representa um número 15 vezes maior do que em 2006. Uma pesquisa feita pela Nielsen em 2017 demonstra que 81% dos consumidores acreditam que as empresas devem colaborar na preservação do meio ambiente e mais de 60% estão preocupados com questões ambientais. Portanto, os fatores ESG tem ganhado mais visibilidade porque são considerados essenciais para a sociedade e visam aumentar oportunidades, mitigando riscos e gerando valor para as empresas (MISHRA, 2020).

No Brasil, o tema ESG cresceu recentemente com a movimentação das empresas em adoção a práticas ESG e comportamento de investidores considerando fatores ESG em decisões de investimento. Entretanto, ainda há muito a ser desenvolvido no mercado brasileiro, por exemplo, a avaliação (*rating*) ESG para as empresas. Enquanto as métricas para avaliar esses fatores estão em estudo no mercado brasileiro, uma oportunidade para as empresas é rever quais iniciativas estão alinhadas às práticas ESG considerando a matriz de materalidade para os fatores, isso significa o nível de relevância por setor inerentes ao negócio (BELINKY, 2021). Afinal diferentes setores do mercado têm diferentes

[1] Who Cares Win

Figura 1 – Fatores ESG e tópicos de sustentabilidade abordado por cada fator



Fonte: A autora (2021)

contribuições para cada fator.

1.1 Objetivos

Este trabalho tem como objetivo desenvolver uma análise exploratória de documentos a partir da extração das principais iniciativas ESG utilizando como técnica a modelagem de tópicos via *Latent Dirichlet Allocation* (LDA).

Os objetivos específicos são:

- Realizar pré-processamento adequado de textos de relatórios ESG e sustentabilidade empresas nacionais;
- Identificar os tópicos mais recorrentes;
- Explorar técnicas de aprendizado de máquina adequadas para extração e classificação de principais iniciativas;
- Utilizar uma abordagem com maior explicabilidade das etapas ao invés de métodos complexos.

1.2 Justificativa

Uma forma de observar a percepção do mercado sobre a aderência às práticas ESG pode ser encontrada nos resultados divulgados em relatórios e notícias que podem demonstrar o nível de alinhamento com as práticas ESG e consequentemente, o reflexo na performance das empresas no mercado de ações. Embora os fatores ESG estão em

destaque no mundo corporativo, a falta de diretrizes adotadas pelas empresas é evidente em relatórios e outras publicações disponibilizadas pelas mesmas ao informar seus *stakeholders*.

Uma abordagem interessante proposta para avaliar o alinhamento com melhores práticas ESG tem sido realizada através do Processamento Natural de Linguagem (AMEND, 2020; RAMAN; BANG; NOUBAKHSH, 2020; SOKOLOV et al., 2021). A partir de relatórios e notícias, é possível extrair informações das principais iniciativas e classificá-las segundo os 3 principais pilares (ambiental, social e governança) e tópicos relacionados. A possibilidade de trabalhar com dados não-estruturados (textos) e extrair *insights* valiosos para tomada de decisão pode contribuir para alavancar o desempenho e reconhecimento das empresas.

Portanto, este trabalho, contempla além de estudos sobre a modelagem probabilística de tópicos, também a realização de uma análise exploratória, usando o LDA para extrair os tópicos a serem analisados. Os modelos probabilísticos de tópicos buscam descobrir estruturas temáticas latentes em coleções de documentos. O LDA, por sua vez, é um modelo bayesiano completo embasado na geração de tópicos como distribuições de Dirichlet, descrevendo um modelo capaz de classificar documentos não conhecidos utilizando informações *a priori* (BLEI; NG; JORDAN, 2003). Dessa forma, os resultados obtidos podem auxiliar na tomada de decisão de negócios das empresas.

1.3 Estrutura do Trabalho

O presente trabalho está estruturado da seguinte forma: o Capítulo 2 apresenta trabalhos relacionados ao tema. No Capítulo 3 é apresentado o referencial teórico. O Capítulo 4 contém os resultados obtidos. Por fim, no Capítulo 5 é apresentada a conclusão e sugestões para trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo contempla uma revisão da literatura sobre aplicações de ferramentas de Ciências de Dados em ESG e outros estudos referentes ao setor financeiro, trazendo uma breve comparação entre as principais contribuições acadêmicas e diferentes abordagens utilizadas.

2.1 Trabalhos Relacionados

O Processamento de Linguagem Natural (PLN) é uma parte da inteligência artificial focada na comunicação entre humanos e máquinas. Enquanto a linguagem humana pode apresentar ambiguidade e imprecisão, as máquinas necessitam de mensagens não-ambiguas e precisas para compreender (FISHER; GARNSEY; HUGHES, 2016). Segundo Liddy (2001): "PLN é uma área de técnicas computacionais para analisar e representar textos que ocorrem naturalmente em um ou mais níveis de análise linguística com o propósito de alcançar o processamento de linguagem semelhante ao humano para uma gama de tarefas ou aplicações".

Sob o aspecto financeiro muitos documentos são divulgados sobre o desempenho corporativo passado e resultados de auditorias. O crescimento no volume de documentos (dados não-estruturados) faz com que o PLN tenha um potencial para melhorar a comunicação em áreas de contabilidade, auditoria e finanças (FISHER; GARNSEY; HUGHES, 2016) por conta de futuras expectativas, políticas e práticas incorporadas pelas demonstrações financeiras. Alguns trabalhos fizeram uma revisão sobre essas aplicações e demonstraram a evolução em diferentes setores do mercado financeiro e futuras expectativas aplicando técnicas de mineração de texto que podem extrair valiosos *insights* para gestores e investidores para conhecer melhor o desempenho corporativo (LEWIS; YOUNG, 2019; FISHER; GARNSEY; HUGHES, 2016; HAGENAU; LIEBMANN; NEUMANN, 2013).

A revisão feita por Fisher, Garnsey e Hughes (2016) cita diversos trabalhos em auditoria e finanças que utilizaram algoritmos de PLN para classificação desde regressão logística, máquina de vetores de suporte (SVM), *k-nearest neighbour* (KNN), *term frequency-inverse document frequency* (TF-IDF). Outro uso foi em detecção e aplicação de fraude através de SVM, *Latent Dirichlet Analysis* (LDA) detectando fraude, falência e análises de comunicados de fiscalização de auditoria e contabilidade. Predição de mercado de ações também foi abordada por meio de PLN em 51 estudos encontrados pelos autores, entre as técnicas encontradas, destaca-se KNN, *Suport Vector Machine* (SVM), LDA, *Latent Semantic Analysis* (LSA). Os autores concluem que SVM é o algoritmo mais utilizado entre os estudos, seguido de agrupamento hierárquico, métodos estatísticos e ponderação LDA.

Hagenau, Liebmann e Neumann (2013) analisaram se a predição do preço no mercado de ações seria aprimorada com informações textuais em notícias financeiras. O estudo demonstra que foi possível a melhoria com métodos de mineração de texto existentes usando recursos mais expressivos para representar texto e empregando feedback de mercado como parte do processo de seleção de recursos. A metodologia empregada considera três aspectos relevantes feitos observados pelos autores: a) o conjunto de dados aplicado no estudo, pois alguns textos podem ser mais fáceis de classificar que outros; b) o processamento de atributos, uma etapa de extrema importância em mineração de textos que pode ser dividida em três passos: extração, seleção e representação de atributos e c) a escolha do algoritmo de aprendizado de máquina, tendo como métrica principal a acurácia para mensuração da performance do modelo de classificação.

Este estudo verificou que as abordagens mais comuns utilizam SVM, Naive Bayes e redes neurais. Apesar da escolha do algoritmo de classificação ser importante, a etapa de seleção de atributos desempenha um papel fundamental no aumento de acurácia do modelo, pois reduz ruídos e limita alguns efeitos de *overfitting*. Entre as técnicas de extração de atributos dos textos os autores aplicaram: uma abordagem por dicionário, para determinar a lista de atributos; *Bag-of-words*, que consiste na extração de palavras relevantes nos textos; N-gram, combinações de N palavras, letras ou sílabas. Após a comparação de diferentes técnicas e modelos, a melhor performance obtida pelos autores foi SVM combinada com N-gram (com 2 palavras) que obteve uma acurácia de 76%.

Atkins, Niranjan e Gerding (2018) exploraram a hipótese de que as informações derivadas das notícias provavelmente têm um efeito maior sobre a característica de segunda ordem da volatilidade do mercado do que sobre os valores dos ativos ou sua direção de movimento. Os autores realizaram um estudo empírico e mostraram evidências que apoiam essa hipótese. Construiram um modelo com LDA, sendo eficaz na redução de recursos de linguagem natural, embora se prove ser computacionalmente caro para treinar. A classificação é alcançada usando um algoritmo de Naive Bayes, que funciona bem, apesar da suposição simplista de independência de atributos. Os resultados obtidos mostraram que a precisão da previsão média para a volatilidade, com a chegada de novas informações, é 56%, enquanto o preço de fechamento do ativo é em torno de 49%. Os autores avaliaram os resultados usando uma variedade de ações e ações índices no mercado dos Estados Unidos da América, usando uma fonte de notícias como entrada e concluíram que os movimentos de volatilidade são mais previsíveis do que movimentos de preços de ativos ao usar notícias financeiras como entrada de aprendizado de máquina e, portanto, podem ser potencialmente explorados na precificação de contratos de derivativos via quantificação da volatilidade.

Outra contribuição importante foi feita por Lewis e Young (2019), sabe-se que economistas e investidores confiam em métricas quantitativas do mercado, porém a

utilização de métricas limita *insights* visto que algumas nuances do desempenho financeiro podem conter os principais aspectos do desempenho em discussões verbais assim como os valores organizacionais podem não estar refletidos nos resultados das demonstrações financeiras em tempo hábil. Portanto, o conteúdo qualitativo presente na linguagem de periódicos de demonstrações financeiras pode ser um grande aliado de informações para participantes do mercado para avaliação, monitoramento e administração.

Os autores discutem e comparam alguns métodos de processamento de atributos e algoritmos de aprendizado de máquina como palavras-chave e contagem, dicionários de atributos, classificação Naive-Bayes, similaridade por cosseno e LDA. Entre os resultados obtidos Naive-Bayes combinado *Bag-of-words* teve uma acurácia de 83%.

Como foi apresentado anteriormente, existe uma vasta aplicação de PLN em trabalhos acadêmicos em diferentes áreas, inclusive no mercado financeiro. Contudo, ainda há uma escassez de estudos sobre o uso de PLN para classificação e avaliação de fatores ESG, especialmente na literatura científica nacional.

Este cenário deve mudar com o destaque que os fatores ESG têm assumido no mundo financeiro e corporativo (RAMAN; BANG; NOUBAKHSH, 2020) . Portanto, a crescente relevância do tema deve atrair novas perspectivas para maior extração de *insights* e métricas derivadas por PLN.

Apesar de não ser uma contribuição científica, Amend (2020) utilizou uma abordagem muito interessante com as principais iniciativas extraídas de relatórios de sustentabilidade a partir de LDA e depois agrupou os tópicos por *k-means* para verificar a distribuição. Para obter uma pontuação para cada fator ESG a partir dos tópicos extraídos e ser uma abordagem *data-driven*, coletou notícias de finanças de uma base de dados GDELT para realizar análise de sentimento e servir como *proxy*. A combinação proposta acaba mitigando problemas como *greenwashing*, prática conhecida na qual a divulgação de ações sustentáveis pela empresa não condizem com a imagem que mercado possui a respeito da sua conduta sobre os eixos ESG. Além disso, foi levantada a conexão entre instituições e organizações interligadas às empresas para avaliar o peso sobre a performance ESG.

Em contribuições acadêmicas, Raman, Bang e Noubakhsh (2020) fizeram uso de rede neurais com modelos pré-treinados de *Transformers* como BERT (*Bidirectional Encoder Representations from Transformers*), XLNet e RoBERTa através de uma classificação de texto semi-supervisionada considerando sentenças extraídas de relatórios de sustentabilidade e classificando-as por grau de relevância seguindo de uma análise de tendência de crescimento dos principais fatores ESG nos últimos 5 anos. Com intuito de observar as tendências em discussões sobre ESG sem prever o comportamento de mercado.

Nugent, Stelea e Leidner (2020) também utilizaram rede neurais com 12 camadas através de BERT, o modelo foi pré-treinado com o vocabulário em inglês da Wikipedia e

BookCorpus totalizando 3300M palavras além de notícias do mercado financeiro da Reuters. A partir do modelo pré-treinado, os autores utilizaram 31.605 notícias envolvendo o tema ESG para extrair 20 categorias identificadas por analistas com formação em finanças e investimentos sustentáveis. Devido à escassez de dados devidamente rotulados na área financeira, os autores também realizaram *data augmentation*, para aumentar a quantidade de dados sem coletar novos exemplos. É uma técnica comum em visão computacional que permite a melhoria da acurácia controlando para não haver *overfitting*. O método utilizado foi *back-translation* atráves de implementação open-source Tensor2Tensor, foi possível aumentar o corpus com a tradução original de inglês para francês e retornando para inglês ao manter a semântica do texto original.

Para comparação os autores também aplicaram SVM, observaram melhor precisão em BERT. Eles demonstram que data augmentation pode mitigar o desafio de criação de modelos com pequeno conjunto de datasets e inclusive aumentar a performance. Contudo, a utilização de notícias do mercado financeiro retorna termos mais comuns aos investidores e menos associados aos fatores ESG, que podem assim revelar tópicos relevantes que posteriormente podem trazer valor às empresas e melhorar seus indicadores no mercado financeiro.

Guo et al. (2020) implementaram uma nova estrutura de *deep learning* para prever a futura volatilidade dos preços das ações. A arquitetura proposta denominada pelos autores ESG2Risk, é composta por um fluxo que consiste desde a extração de notícias sobre ESG, transformação por text embedding e análise de sentimento seguido de aplicação de modelo de *deep learning* para predição.

Serafeim (2021) utilizou análise de sentimento para avaliar a performance de empresas em relação aos fatores ESG usando como *proxy* a pontuação de MSCI para ESG. Uma limitação do estudo deve-se ao fato de obter uma avaliação sem realmente detectar os pontos de melhoria que cada empresa pode ter melhor performance a respeito de iniciativas ESG.

Por outro lado, Sokolov et al. (2021) utilizaram o Transformer BERT para classificar 10 tópicos extraídos de notícias do Twitter e depois agrupa-los. Apesar de ser uma abordagem interessante alguns impedimentos levantados pelos autores são a depêndencia de ajuste manual por conta dos limiares para cada categoria ESG e também a susceptibilidade a ruídos, por exemplo, uma campanha de marketing pode melhorar a performance geral de pontuações ESG.

Kiesel e Lücke (2019) aplicaram o método LDA, a fim de determinar ESG nas classificações de crédito. Pois o método é preferencialmente aplicável para a natureza qualitativa e não financeira do ESG. A ferramenta de modelagem de tópicos foi empregada para calcular uma medida ESG para um conjunto de 3.719 relatórios de ação de *rating* entre 2004 e 2015 publicados pelo Moody's Investors Service. Para analisar as reações do

mercado de capitais para as classificações sob efeito dos fatores ESG, o estudo utiliza uma regressão multivariada para ações e *Credit Default Swap* (CDS) no dia em que relatório é publicado. Os autores encontraram uma pequena influência da determinação de rating dos fatores ESG, no qual o aspecto mais relevante foi de Governança Corporativa. Além disso, descobriram que todos os critérios ESG são relevantes para capital e dívida de investidores e demais *stakeholders*.

O quadro comparativo abaixo (Tabela 3) resume os métodos, algoritmos e medidas de performance utilizadas pelos estudos discutidos acima. Pode-se observar diferentes abordagens com relação ao tema ESG e a predominância do uso de *deep learning* em alguns trabalhos. As abordagens utilizadas também possuem diferentes finalidades, como o presente estudo focou na análise exploratória e modelagem de tópicos das principais iniciativas difundidas no mercado brasileiro obter maior explicabilidade e por fim, reprodutibilidade nos métodos utilizados mais relevantes para definição da metodologia. Assim como, Lewis e Young (2019) enfatizam o problema de usar abordagens com algoritmos "caixa preta" pela falta de transparência ao descrever as etapas e escolhas envolvidas na aplicação dessas técnicas, optar por métodos mais simples e que apresentam explicações sobre quais parâmetros garantem maior entendimento dos processos utilizados. Por exemplo, como *stop-words* foram eliminadas ou como e porquê *stemming* foi utilizado. Por isso, levando em consideração outros trabalhos de PLN voltados ao setor financeiro é possível combinar outras abordagens mais transparentes e menos complexas para avaliar a performance em classificação de textos.

Tabela 1 – Sumário de estudos relacionados

Autores	**Tipos de dados**	**Método**	**Mineração de Texto**	**Precisão**	**F1-score**	**Revocação**	**Acurácia**
Amend, A. (2020)	Notícias e relatórios	Modelo de tópico	LDA, tf-idf	-	-	-	-
Raman; Bang; Nourbakhsh (2020)	Relatórios	BERT, RoBERTa e XLNet	-	-	78,3%	-	78,1%
Nugent; Stelea; Leidner (2020)	Notícias	BERT; SVM	*tf-idf*	84% ; 76%	84%; 75%	84%; 75%	-
Guo et al. (2020)	Notícias	BERT	-	-	-	-	-
Serafeim (2020)	Dados MSCI e valores ESG	Análise de sentimento	-	-	-	-	-
Sokolov et al. (2020)	Dados MSCI e notícias	BERT	-	60%	-	60%	-
Kiesel e Lücke (2019)	Relatórios de rating	Regressão multivariada	LDA	-	-	-	-
Atkins; Niranjan; Gerding (2018)	Notícias	Näive Bayes	LDA	63%	60,3%	61,5%	61,5%
Lewis e Young (2019)	Notícias	Näive Bayes	*Bag-of-words*	-	-	-	83%
Hagenoau; Liebmann; Neumann (2013)	Notícias	SVM	N-gram	-	-	-	76%

Fonte: A autora (2021).

3 METODOLOGIA

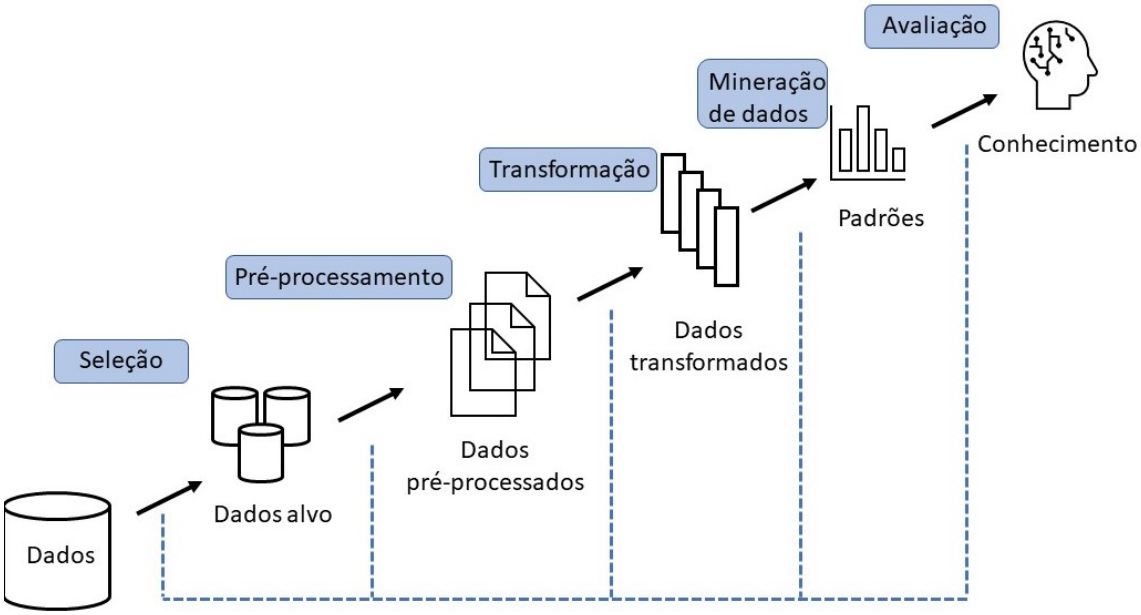
3.1 Considerações gerais

Esta seção visa a apresentação da literatura científica sobre fundamentos de Processamento de Linguagem Natural (PLN) assim como técnicas aplicadas para o desenvolvimento deste trabalho. Por ser um tema muito recente há poucos trabalhos referentes às técnicas de mineração e classificação de texto aplicadas aos fatores ESG. Por fim, a revisão seguinte traz fundamentos no estado da arte de técnicas e métodos que podem ser aplicados ao tema.

Descoberta de Conhecimento e Mineração de Texto

Devido ao crescente volume de dados disponibilizados de forma estruturada ou não-estruturada, teorias e ferramentas computacionais surgiram para a extração automática de conhecimento de bases de dados. Este campo de pesquisa é conhecido como *Knowledge Discovery in Databases* (KDD). Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), o KDD é o processo de identificação de padrões singulares, potencialmente úteis e compreensíveis embutidos nos dados, que consistem em diversas etapas (Figura 2).

Figura 2 – Etapas do processo de Descoberta de Conhecimento em Banco de Dados (KDD)



Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996)

Conforme representado na Figura 2, estas etapas são constituídas dos seguintes processos:

1. Desenvolvimento do entendimento da aplicação: define a meta do processo KDD visando o objetivo a ser alcançado.

2. Criação de dados alvo: consiste na seleção do dataset, ou subset de variáveis ou amostras.

3. Pré-processamento e limpeza de dados: operações básicas para redução de ruídos, decisão de estratégias para dados nulos ou desbalanceados.

4. Redução de dados e projeção: uso de redução de dimensionalidade dependendo do objetivo da tarefa para selecionar variáveis de interesse.

5. Escolha do algoritmo de mineração de dados: seleção do método aplicado para a busca de padrões nos dados (classificação, clusterização, regressão, sumarização, etc).

6. Interpretação: esta etapa consiste na visualização e interpretação dos dados extraídos dos modelos.

7. Consolidação do conhecimento adquirido: incorporação e implantação do conhecimento assim como sua documentação.

Este processo também pode ser aplicado em informações textuais. Enquanto banco de dados armazenam registros estruturados, técnicas específicas devem ser aplicadas para possibilitar a descoberta de informações implícitas sobretudo na etapa de pré-processamento. Tais técnicas são apresentadas a seguir.

3.2 Seleção dos Dados

O conjunto de dados utilizado para a presente pesquisa foi coletado em diversas fontes disponíveis abertamente em sites de empresas nacionais que reportam relatórios de sustentabilidade ou ESG em websites. Com intuito de trazer representatividade para o estudo, selecionou-se documentos de diferentes setores.

3.3 Pré-Processamento

Tokenização

É a tarefa de segmentação do texto em partes chamadas *tokens*, que significa uma instância de uma sequência de caracteres de um documento que possui um significado semântico para o processamento (MANNING; RAGHAVAN; SCHÜTZE, 2009).

Expressão regular

Expressões regulares são padrões encontrados no texto que servem para criação de regras para extração de informações ou limpeza dos dados (VAJJALA et al., 2020). Geralmente utiliza-se expressões regulares no processo de tokenização para padronizar os *tokens*, tal como, transformar todo o documento em letras minúsculas e com remoção da pontuação. Tendo em vista que cada *token* é composto de uma sequência de caracteres diferentes, isso implica em possíveis repetições desnecessárias para a frequência de uma única palavra. Por isso, surge a necessidade de padronização do documento, e alguns comandos para expressões regulares se encontram inseridos no pacote `NLTK`[1], por exemplo.

Stop-words

São termos com alta frequência no texto porém com pouca relevância para o contexto do documento e normalmente são removidos para melhorar o desempenho do processamento (MANNING; RAGHAVAN; SCHÜTZE, 2009). Podem ser artigos, preposições, conjunções, etc. Para fazer a remoção dessas palavras é necessário criar uma *stop list* com estes termos e filtrar do texto analisado. Alguns pacotes, como o pacote `NLTK` contém uma lista de stop-words para alguns idiomas que facilita esta etapa de pré-processamento. Entre algumas palavras implementadas no pacote, temos: “a”, “ao”, “aos”, “aquela”, “aquelas”, “aquele”, “aqueles”, “aquilo”, “as”, “e”, “para”, “ou”, etc.

Lematização e Stemming

A lematização é uma técnica para transformar toda palavra para a forma canônica levando em consideração a classe gramatical, no caso de verbos, para a forma infinitiva e substantivos e adjetivos, para a forma masculina singular. Por exemplo: “propõem”, “propondo” e “propuseram” pertencem ao verbo “propor”, e portanto, seria o lema retornado.

Enquanto a radicalização ou *stemming* é a técnica de redução para o radical da palavra, que diminui as variações morfológicas, removendo sufixos e prefixos. No caso do verbo “propor” a redução seria “prop”, considerando as flexões do verbo.

A meta de ambas as técnicas é a redução de forma flexional da base de uma palavra. Contudo, *stemming* geralmente é um processo heurístico que elimina sufixos de palavras para atingir o objetivo, enquanto a lematização considera análise morfológica de palavras, normalmente com o objetivo de remover apenas as extremidades flexionais e retornar a base ou a forma de dicionário de uma palavra, que é conhecida como lema (MANNING; RAGHAVAN; SCHÜTZE, 2009).

[1] https://www.nltk.org/

3.4 Transformação e mineração de dados

Esta etapa consiste na extração e representação de features em uma matriz valor-atributo para que os algoritmos de classificação apresentadas na seção seguinte possam processar a informação contida em cada documento.

Bag-of-Words

Uma técnica de representação de texto em PLN é *Bag-of-Words*, principalmente em problemas de classificação de texto. A ideia principal consiste em representar o texto como uma coleção de palavras ignorando a ordem e o contexto, assumindo que o texto pertence a uma determinada classe caracterizada por um conjunto de palavras. Portanto, se dois textos possuem praticamente as mesmas palavras, eles pertencem a mesma classe. A técnica *Bag-of-Words* é semelhante ao *one-hot encoding*, no qual mapeia palavras para IDs inteiros exclusivos entre 1 e $|V|$, onde V é o conjunto do vocabulário do corpus. Cada documento do corpus é então convertido em um vetor de $|V|$ dimensões onde na i-ésima componente do vetor, $i = wid$, é simplesmente o número de vezes que a palavra w ocorre no documento, ou seja, simplesmente pontuamos cada palavra em V por sua contagem de ocorrências no documento. Algumas vantagens dessa técnica são: a facilidade implementação e entendimento; documentos com o mesmo uso de vocabulário têm suas representações vetoriais mais próximas entre si no espaço euclidiano, capturando a semelhança semântica dos documentos. Contudo entre as desvantagens, destaca-se: o aumento do vetor de acordo com o tamanho do vocabulário, precisando limitar a um número n de palavras mais frequentes; não associa a semelhança entre palavras diferentes, mas com mesmo significado e por fim, as informações sobre a ordem das palavras se perdem na representação (VAJJALA et al., 2020).

Term Frequency – Inverse Document Frequency

O algoritmo TF-IDF (*Term Frequency – Inverse Document Frequency*) representa a frequência de termos, isso significa a frequência da palavra t no documento d, por exemplo. Serve para balancear a importância das palavras no texto, pois uma palavra onipresente pode perder sua relevância diante de palavras que ocorrem em conjunto em poucas vezes no texto ou em uma coleção (VAJJALA et al., 2020). A expressão tf é dada pela frequência do termo t:

$$\mathrm{tf}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

onde $f_{t,d}$ é a contagem bruta de um termo, ou seja, o número de vezes que esse termo t ocorre no documento d. Existem várias outras maneiras de definir a frequência do termo Enquanto o termo IDF é dado por:

$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|},$$

onde N denota o número total de documentos no corpus, isto é, $N = |D|$. O conjunto $\{d \in D : t \in d\}$ representa o subconjunto de documentos de D em que o termo t aparece (ou seja, $\text{tf}(t, d) \neq 0$). Se o termo não estiver no corpus, isso levará a uma divisão por zero. Portanto, é comum ajustar o denominador para $1 + |\{d \in D : t \in d\}|$.

O produto dos dois termos gera a pontuação TF-IDF. Há outras variações da fórmula utilizada para TF-IDF. Assim como, *Bag-of-Words*, os vetores TF-IDF são utilizados para calcular a similaridade entre textos por distância Euclidiana ou similaridade por cosseno. Portanto, é comumente utilizada para classificação de textos, porém sofre com a maldição da dimensionalidade (VAJJALA et al., 2020).

3.5 Modelagem e avaliação

Com a crescente quantidade de informação textual existente, se torna humanamente impossível absorver ou captar informações relevantes. Logo, técnicas como a mineração de textos tem colaborado bastante para a obtenção de dados. A mineração de texto utiliza várias técnicas avançadas de mineração de dados, aprendizado de máquinas, recuperação de informação, extração de informação, linguística computacional e Processamento de Linguagem Natural VAJJALA et al. (2020). Logo, é relevante o desenvolvimento de maneiras automáticas para a extração de informações em textos, pois envolve a manipulação de dados não estruturados, sendo esta uma tarefa desafiadora.

Deste modo, pesquisadores de aprendizado de máquinas desenvolveram a modelagem probabilística de tópicos, que se trata de um conjunto de algoritmos com o objetivo de obter informações temáticas em grandes arquivos de texto. Segundo BLEI, NG e JORDAN (2003), estes algoritmos são métodos estatísticos que analisam as palavras dos textos originais, visando descobrir os temas abordados, como eles se conectam e como mudam ao longo do tempo.

Nesta etapa, detalha-se a abordagem que será utilizada a busca de padrões nos dados com o intuito de extrair os principais tópicos encontrados nos documentos.

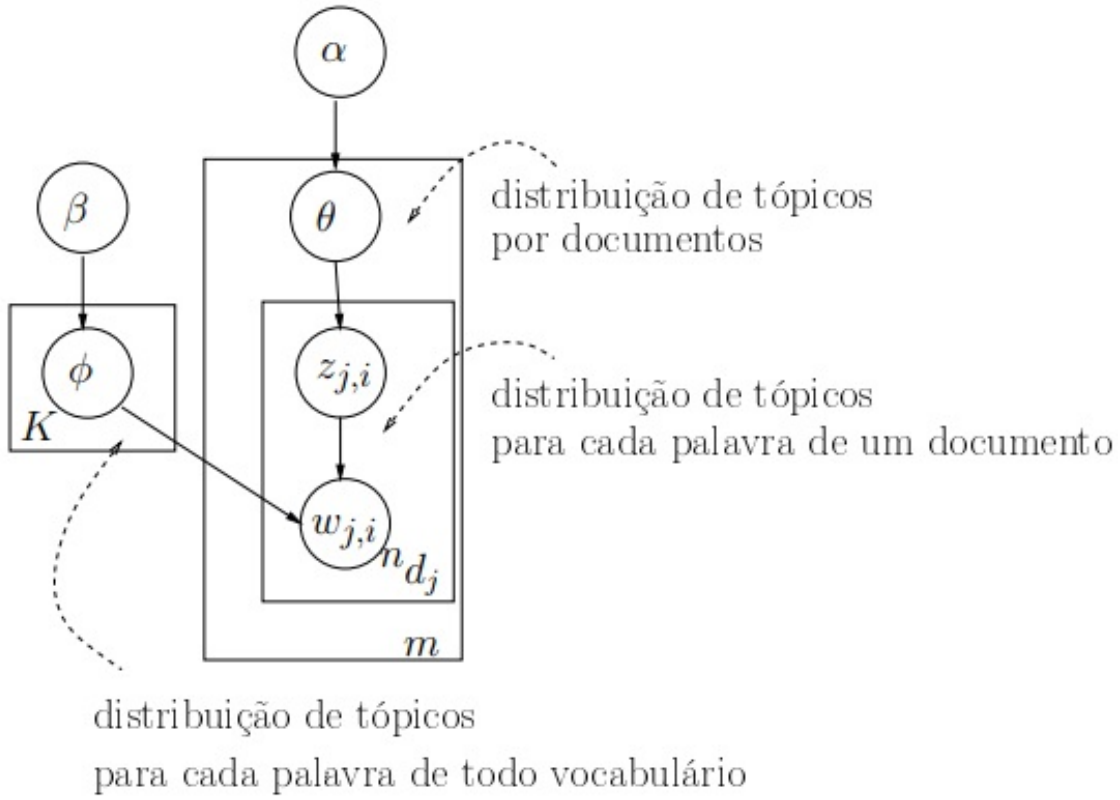
Latent Dirichlet Allocation

Um modelo probabilístico generativo e aplicado para classificação de documentos proposto por BLEI, NG e JORDAN (2003) é o *Latent Dirichlet Allocation* (LDA) reconhecido na literatura como estado da arte para modelos probabilísticos de tópicos. A distribuição de *Dirichlet* é utilizada para amostrar a distribuição de tópicos. O resultado da amostragem é usado para alocar as palavras de diferentes tópicos e compõem os docu-

mentos. Portanto, o significado do nome *Latent Dirichlet Allocation* expressa o propósito do modelo alocar os tópicos latentes que encontrados conforme a distribuição de *Dirichlet* (FALEIROS, 2016).

A Figura 3 representa de forma gráfica o processo generativo por meio de uma rede Bayesiana, onde cada vértice corresponde a uma variável e cada aresta a uma relação de dependência. Os hiper-parâmetros α e β estão no nível de coleção e determinam o comportamento dos tópicos. Quando o valor de α for alto, os documentos provavelmente compreenderão uma maior mistura de tópicos, e no caso contrário, a mistura será de poucos tópicos. Por outro lado, se o valor de β for alto, cada tópico terá uma maior probabilidade de possuir misturas de várias palavras, senão, o tópico será formado por poucas palavras.

Figura 3 – Representação gráfica do modelo LDA.



Fonte: Faleiros (2016)

Conforme Faleiros (2016), os itens apresentados na figura acima representam:

- K - número de tópicos;
- n - número de palavras do vocabulário;
- m - número de documentos;
- n_{d_j} - número de palavras em um documento d_j, onde $1 \leq j \leq m$;

- θ - distribuição de tópicos por documentos;
- ϕ - distribuição dos tópicos sobre as palavras do vocabulário;
- θ_j - vetor com a proporção dos tópicos para o documento d_j, onde $1 \leq j \leq m$;
- ϕ_k - vetor com a proporção das palavras do vocabulário para o tópico k, onde $1 \leq k \leq K$;
- α - Priore da distribuição de Dirichlet, relacionada a distribuição documento-termo;
- β - Priore da distribuição de Dirichlet, relacionada a distribuição tópico-palavra;
- w_i - i-ésima palavra do vocabulário, onde $1 \leq i \leq n$;
- $w_{j,i}$ - palavra w_i observada no documento d_j, onde $1 \leq j \leq m$ e $1 \leq i \leq n$;
- $z_{j,i}$ distribuição de tópicos associado a palavra $w_{j,i}$ no documento d_j, onde $1 \leq j \leq m$ e $1 \leq i \leq n$.

O método fornece, em comparação com as abordagens quantitativas, uma vantagem especialmente para a análise de aspectos qualitativos. Assim, a ideia básica do LDA é que os documentos são misturas aleatórias de vários tópicos latentes ou desconhecidos, onde cada tópico é classificado por uma distribuição de palavras (BLEI; NG; JORDAN, 2003). O modelo difere de outros métodos de modelagem de tópico já que pode processar uma amostra de dados nunca vista anteriormente e não há limitações por meio de um conjunto de dados de treinamento (BELLSTAM; BHAGAT; COOKSON, 2017; BLEI; NG; JORDAN, 2003).

Além disso, a concepção de LDA permite que os documentos dentro de uma amostra possam ser associados a vários tópicos, enquanto os tópicos não precisam ser predefinidos além do seu número K (BLEI; NG; JORDAN, 2003). Além da simplicidade, isso ainda explica a possibilidade de que as palavras tenham significados diferentes em relação ao contexto e possam ser alocadas em vários tópicos. Isso também representa uma vantagem inerente em comparação com procedimentos padrão de lista de palavras, pois permitem a utilização de palavras apenas para um tópico.

Para a aplicação do modelo LDA, é necessário especificar o número total de tópicos K que irá produzir dois resultados do corpus:

1. a distribuição das frequências de palavras para cada tópico k presentes nos documentos;
2. a distribuição dos tópicos nos documentos. Por exemplo, as frequências nas quais os tópicos são usados nos documentos.

4 RESULTADOS

Esta seção contempla o detalhamento das atividades desenvolvidas do projeto para alcançar o objetivo proposto, detalhando as tarefas realizadas até os resultados obtidos conforme a etapas demonstradas na Figura 4.

Figura 4 – Etapas do estudo desenvolvido.

Fonte: A autora (2022).

4.1 Coleta para base de Dados

O conjunto de dados utilizado para a presente pesquisa foi coletado em diversas fontes disponíveis publicamente em sites de empresas nacionais que reportam relatórios de sustentabilidade ou ESG em websites no formato PDF. Com intuito de trazer representatividade para o estudo, selecionou-se documentos de diferentes setores. No total, foram utilizados 10 relatórios de sustentabilidade e ESG de empresas relacionadas aos seguintes setores: uma do setor de produção de alimentos e bebidas (ORG-1),uma companhia aérea (ORG-2), uma empresa de varejo (ORG-9), duas instituições financeiras (ORG-3 e ORG-4), cinco do setor industrial sendo duas empresas da área de mineração (ORG-5 e ORG-8), uma da área de cosméticos (ORG-6), uma do setor de combustíveis (ORG-7) por fim, uma companhia de telecomunicações (ORG-10).

4.2 Pré-processamento e análise descritiva

Na etapa de pré-processamento foi realizada a transformação dos relatórios em arquivos na extensão *txt* através do pacote `pdftotext`[1]. Observou-se a ausência de padronização nos documentos desde o tamanho ou vocabulário utilizado. Então para o tratamento inicial foi necessário realizar a tokenização de todos os arquivos com o pacote `NLTK`[2] pelo método ".wordtokenize" por se tratar de dados não-estruturados. O pacote `NLTK` possui diversas ferramentas para pré-processamento de textos e mostrou-se eficiente para desempenhar as tarefas seguintes.

Por conter diversas expressões que podem dificultar o processamento de textos, foi realizado um intenso trabalho de REGEX para eliminar e padronizar palavras relevantes para a análise. Assim, foram removidos caracteres desconhecidos, espaços duplos, aspas, quebra de sílabas por hífen (para formar uma única palavras) e endereços de páginas da internet (URLs).

Em textos encontramos diversas palavras com alta frequência que não adicionam uma relevância para o contexto ou tópico a ser explorado, são conhecidas como *stop-words*. O pacote `NLTK` possui uma coleção de palavras definidas como as principais *stop-words* onde constam 204 palavras na língua portuguesa, foi então aplicado esse método através de uma função em todos os dados onde realiza a transformação do texto em palavras minúsculas, remoção de pontuação e transformação das palavras em *tokens*. A tabela seguinte (Tabela 2) apresenta a quantidade de palavras contidas nos documentos originais e pré-processados. O código utilizado encontra-se no `Github`[3].

Tabela 2 – Quantitativo da base de dados.

Organização	Quantidade de tokens	Quantidade de tokens sem stopwords
ORG-1	41158	16664
ORG-2	23425	10700
ORG-3	89973	41723
ORG-4	180241	85348
ORG-5	23716	11672
ORG-6	82094	38232
ORG-7	132505	62376
ORG-8	66775	32369
ORG-9	24920	11744
ORG-10	6652	3105
Total	**671459**	**313933**

Para analisar as palavras mais frequentes encontradas no conjunto de dados, foi feita nuvem de palavras (Figura 5). É possível observar a predominância de algumas

[1] https://pypi.org/project/pdftotext/
[2] https://www.nltk.org/
[3] https://github.com/karen-lopes/NLP$_{E}SG_{i}niciatives$

palavras comuns em todos os textos. A utilização da nuvem de palavras foi fundamental para encontrar essas palavras dominantes e extrair outras *stop-words* encontradas no conjunto de dados, mas também, foi capaz de demonstrar os termos mais utilizados e específicos em cada texto. Na Figura 5 as palavras 'crédito', 'risco', 'desempenho financeiro' e 'compliance' remetem às instituições financeiras dos relatórios das ORG-3 e ORG-4. Enquanto as palavras 'lixo' e 'Amazônia' se destacam dos outros documentos, pois neste documento referente ao relatório da ORG-6 há o foco do negócio para a região e também acompanhado da preocupação dos resíduos gerados pela indústria.

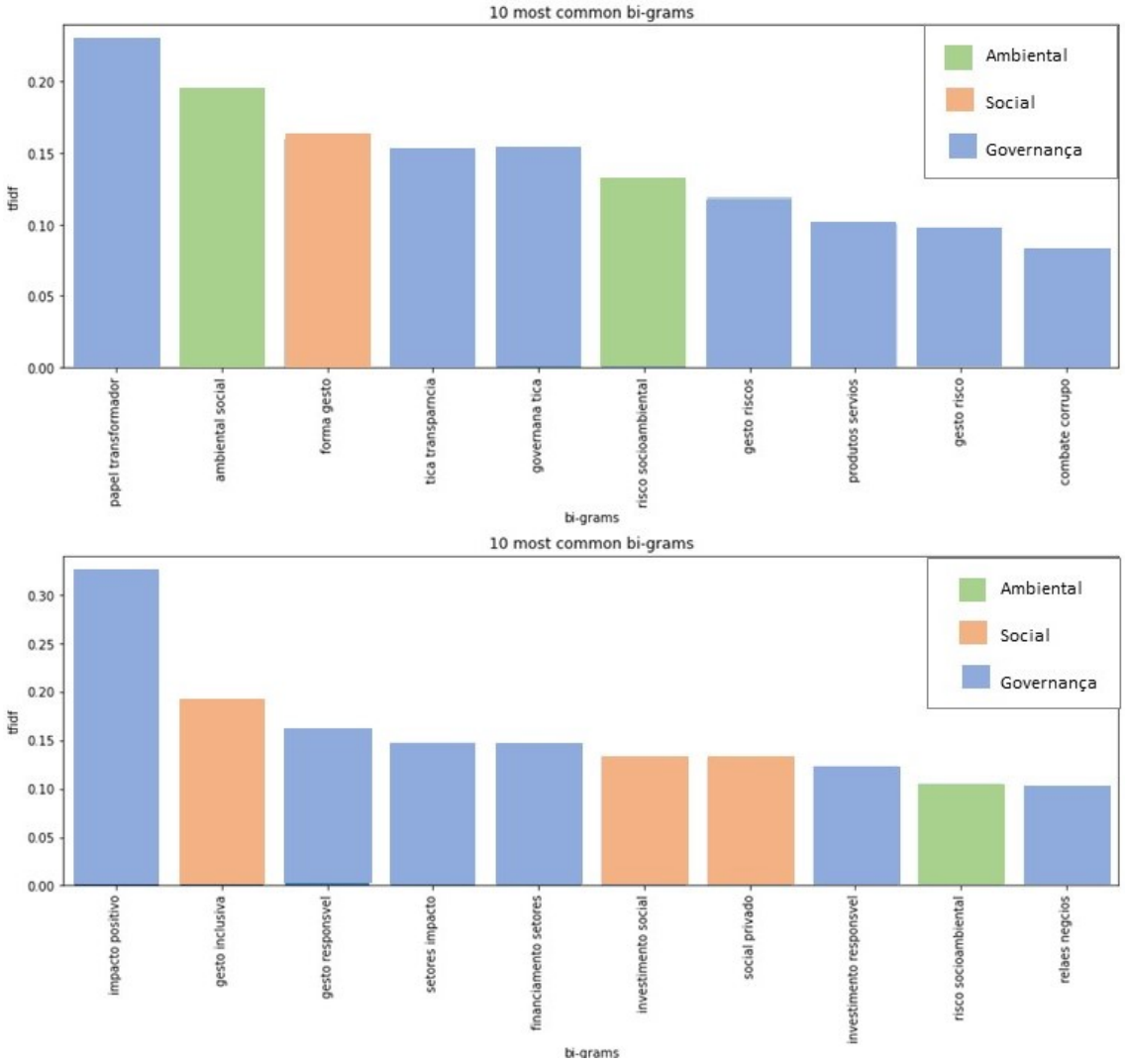
Figura 5 – Nuvem de palavras registradas na base de dados por documento.

Fonte: A autora (2022).

A normalização realizada foi através da técnica de radicalização (*stemming*). Pois como discutido anteriormente, a técnica de *stemming* opera em uma única palavra sem conhecimento do contexto e, portanto, não pode discriminar entre palavras que têm significados diferentes dependendo da classe gramatical. Além disso, são normalmente mais fáceis de implementar e executar com mais rapidez. Após as etapas de pré-processamento, houve a necessidade de salvar os arquivos pré-processados em formato *txt*, compondo então a coleção de documentos a ser analisada. A primeira etapa a ser criada foi a *Bag-of-words*, para avaliar a frequência de ocorrência em espaço vetorial.

Assim como a ponderação de termos por TF-IDF para avaliar a importância de uma palavra de um documento em relação ao *corpus*. Com o uso de TF-IDF, foi possível realizar uma análise de bigramas que reflete as expressões comumente utilizadas nos documentos e intrínsecas ao setor de negócio. Por exemplo, como em instituições financeiras (ORG-3 e ORG-4) são encontradas expressões relacionadas à gestão de riscos, inclusiva ou responsável e outros fatores como combate à corrupção onde naturalmente o pilar de Governança tem maior contribuição com o negócio (Figura 6).

Figura 6 – Bigramas encontrados na base de dados em empresas do setor financeiro com predominância de pilares de Governança.

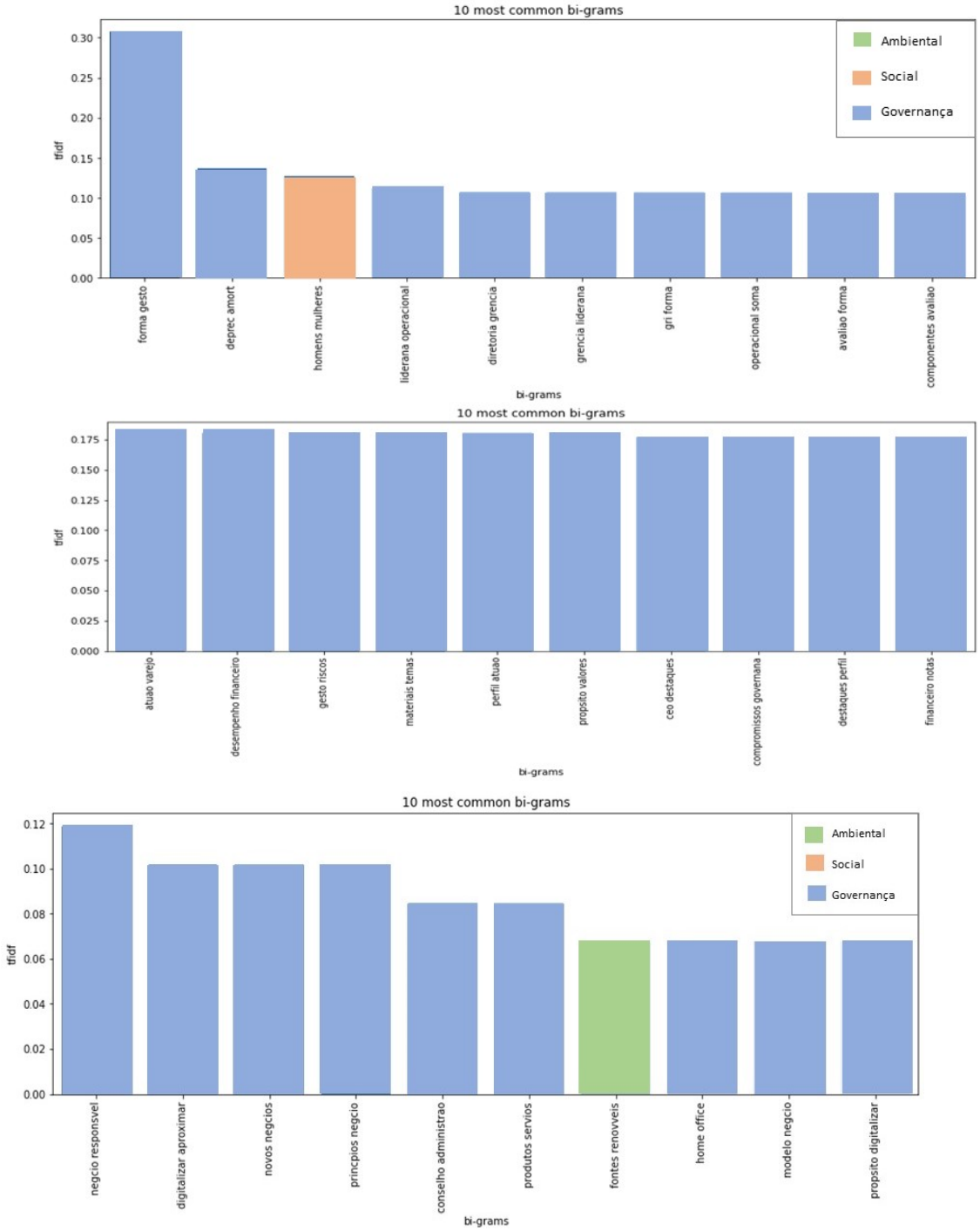


Fonte: A autora (2022).

Assim como em instituições financeiras, foi observado a predominância de aspectos relacionados à Governança nas empresas de alimentos (ORG-1), varejo (ORG-9) e telecomunicações analisadas (ORG-10) como pode ser observado na Figura 7. Com temas muito relacionados à gestão de riscos, resultados financeiros, gerência, forma operacional, etc. Os bigramas encontrados são condizentes com a natureza do negócio, mas é importante

ressaltar explorar outros aspectos dos pilares ambientais ou sociais podem trazer maiores contribuições para a sociedade.

Figura 7 – Bigramas encontrados na base de dados em diferentes empresas do setor alimentício, varejo e telecomunicações, respectivamente, com predominância de pilares de Governança.

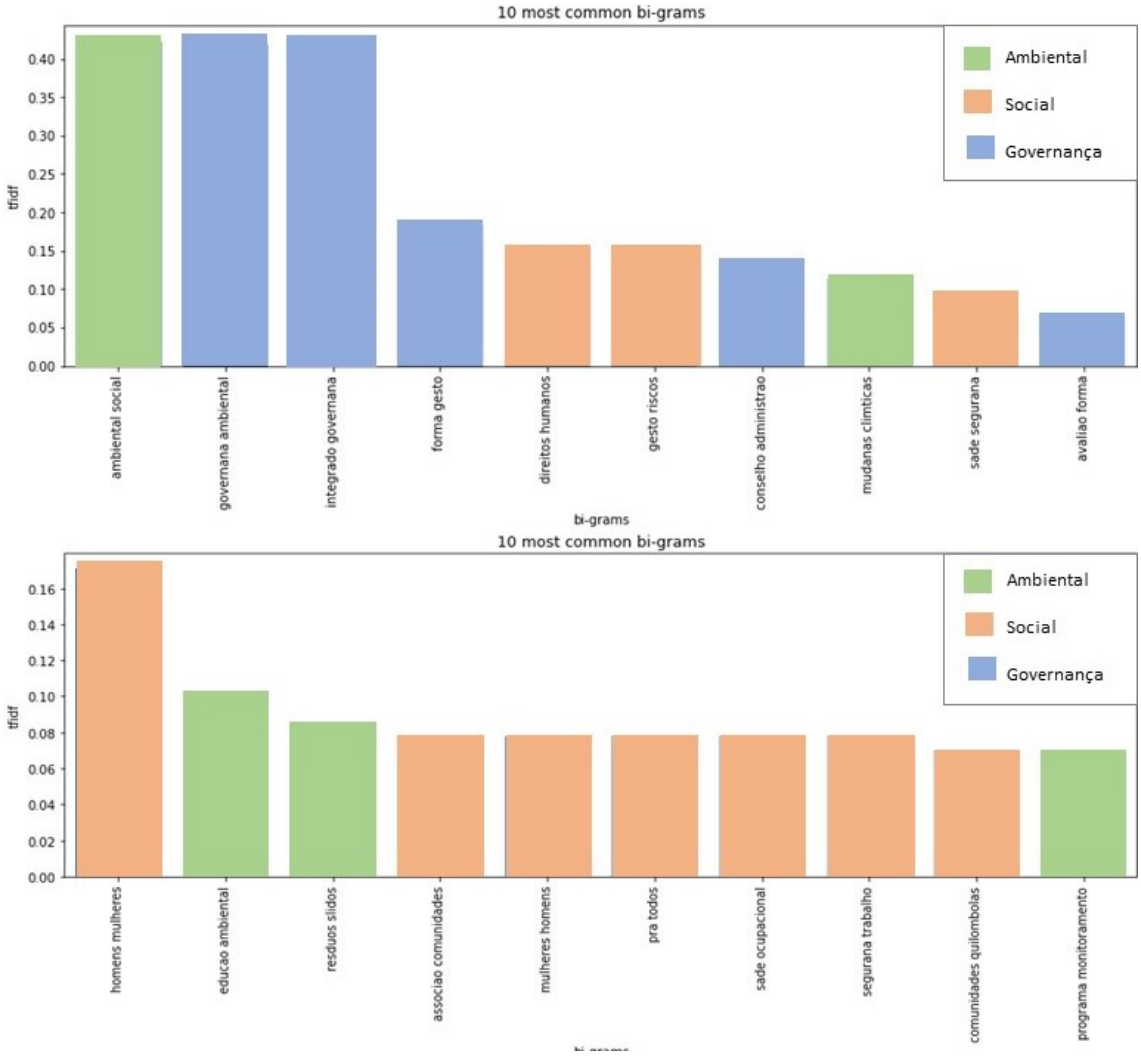


Fonte: A autora (2022).

Em contrapartida, empresas do setor industrial como mineração (ORG-5 e ORG-8), pode-se observar maior predominância de bigramas relacionados aos aspectos ambientais e

sociais (Figura 8). Alguns exemplos são: resíduos sólidos, mudanças climáticas e gestão de riscos, saúde, segurança, comunidade e direitos humanos sob a perspectiva social. É coerente com a atuação do negócio, pois há uma maior cobrança pela interferência no meio ambiente onde essas empresas atuam e por isso é necessário garantir menos impactos ambientais, mas também, a relevância do aspecto social para promover o bem-estar de comunidades e segurança para seus empregados é uma prioridade neste setor.

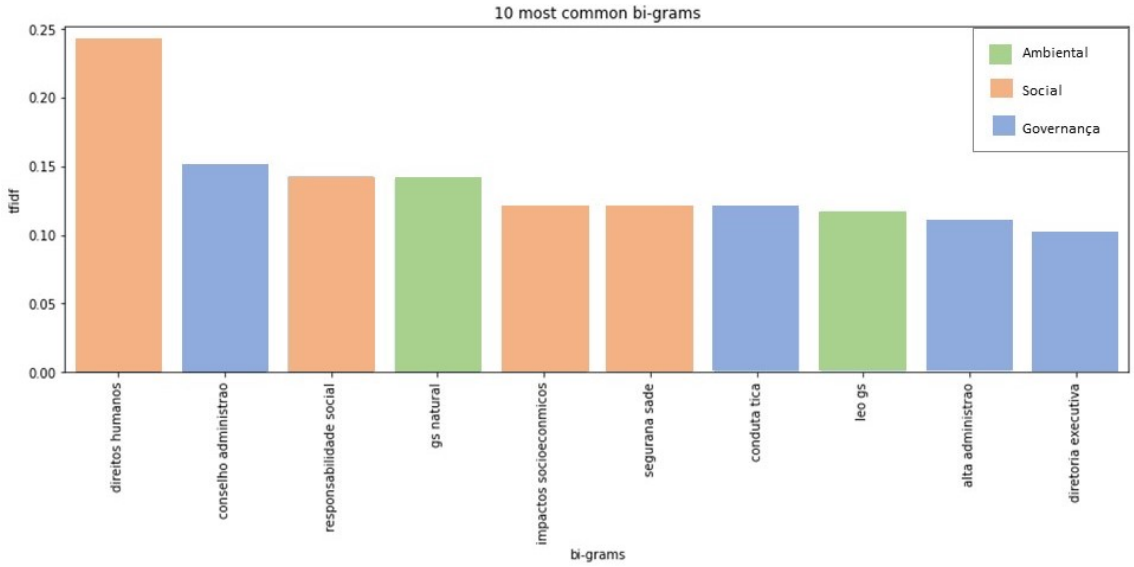
Figura 8 – Bigramas encontrados em relatórios de empresas do setor mineral apresentam predominância de expressões dos pilares sociais e ambientais.



Fonte: A autora (2022).

Ainda a respeito de empresa do setor industrial, para a empresa do setor de combustíveis (ORG-7) há uma certa distribuição mais equilibrada para os três pilares ESG (Figura 9), sendo o maior destaque para 'direitos humanos' que se enquadra no fator Social e o documento reitera o compromisso de respeito aos direitos humanos com base nos princípios do Pacto Global das Nações Unidas e inspirado nos Objetivos de Desenvolvimento Sustentável, por isso o bigrama encontrado reflete essa relevância do tema para o negócio.

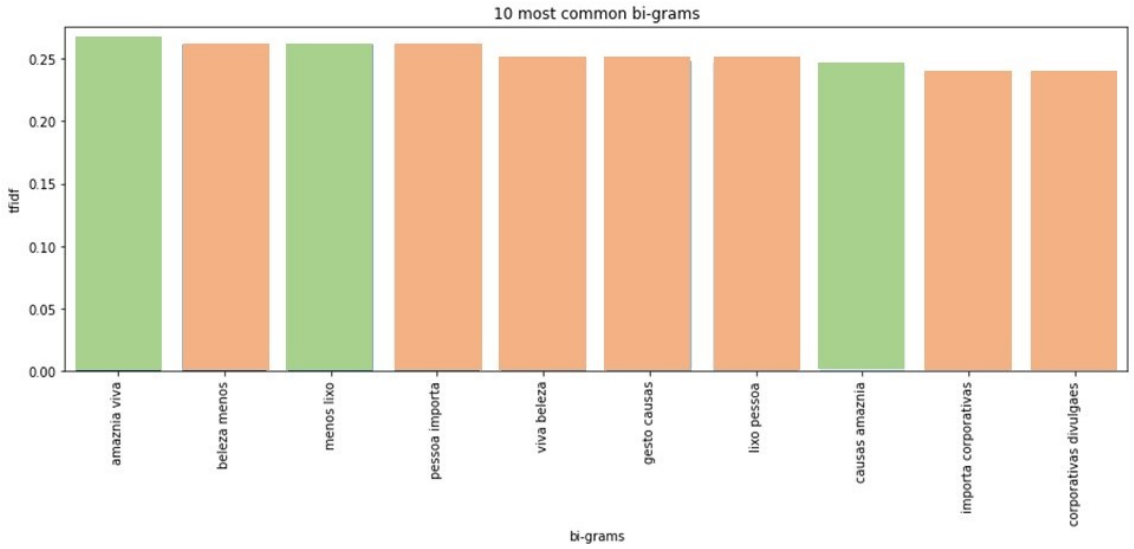
Figura 9 – Bigramas encontrados em relatórios de empresas do setor de combustíveis apresenta distribuição mais equilibrada para os pilares ESG.



Fonte: A autora (2022).

Para a empresa de cosméticos (ORG-6), é possível observar causas relacionadas a causas ambientais e sociais, em especial à preservação da Amazônia devido a atuação da empresa na região Amazônica e também a importância dessa região diante do contexto global (Figura 10).

Figura 10 – Bigramas encontrados em relatórios de empresas do setor mineral apresentam predominância de expressões dos pilares sociais e ambientais.

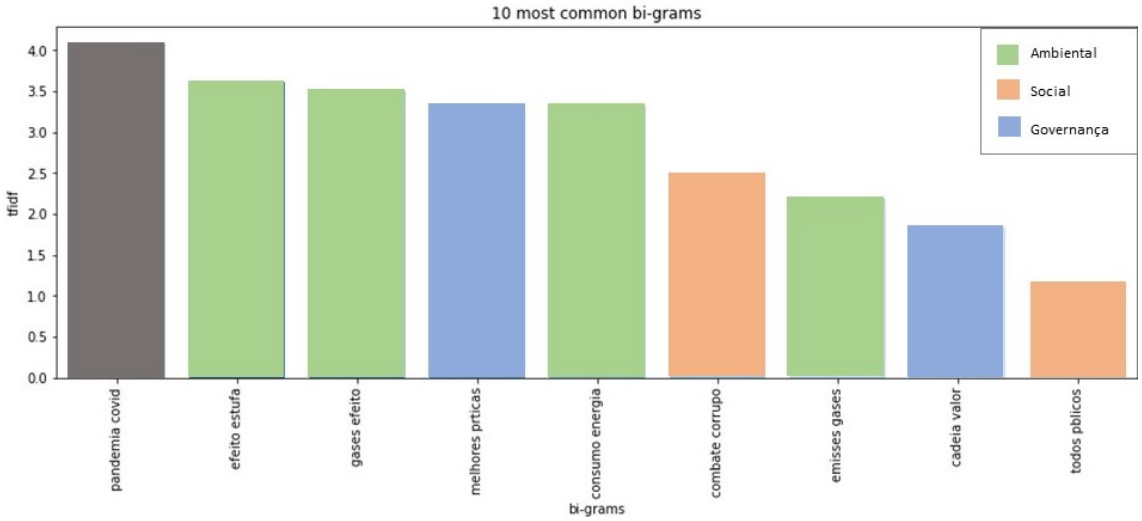


Fonte: A autora (2022).

Ao analisar bigramas para todos os documentos, é importante ressaltar que os bigramas mais frequentes, o maior destaque é para a pandemia de Covid-19 (Figura 11), pois todos os relatórios analisados são do ano de 2020 (início da pandemia) e por isso

tratam do assunto. Além disso, bigramas relacionados a emissões de gases do efeito estufa são frequentes também, provavelmente pelo destaque no ano sobre queimadas na Amazônia e o impacto na emissão de gases do efeito estufa[4].

Figura 11 – Bigramas encontrados em toda base de dados.



Fonte: A autora (2022).

4.3 Mineração de texto

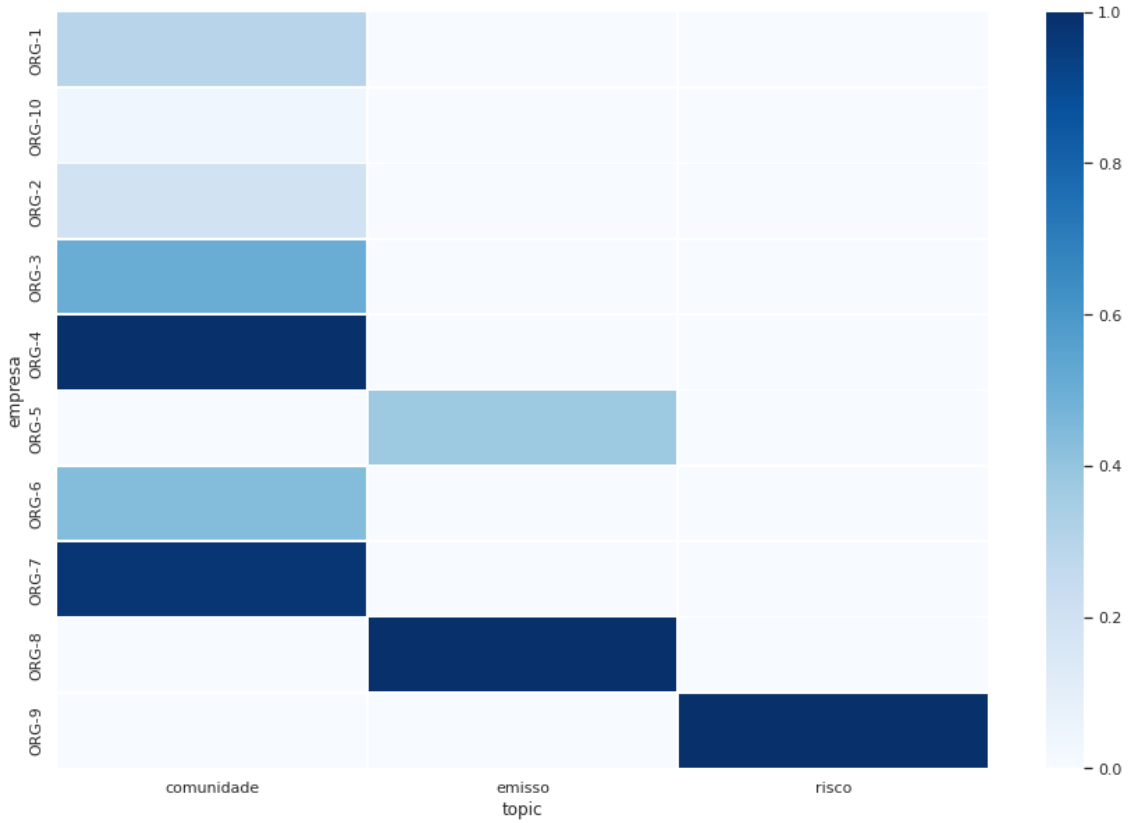
Com a base de dados pré-processado, foi aplicado o modelo para extração dos tópicos. Para gerar o dicionário e corpus utilizou-se a biblioteca `Gensim`[5]. O dicionário possui o registro das palavras contidas na base de dados, atribuindo um identificador único, e também um contador para as ocorrências de tal palavra na base de dados. O corpus, por sua vez, é uma combinação de todos os documentos de texto, sendo composto por uma representação matricial entre documentos e termos. O corpus é usado pelo LDA para procurar padrões na matriz documento-termo.

O mapa de calor apresenta a intensidade da relação entre as empresas e três principais iniciativas encontradas, é representado abaixo (Figura 12). Pelo gráfico a maioria dos documentos (ORG-1, ORG-2, ORG-3, ORG-4, ORG-6) reportam a palavra 'comunidade', logo estando relacionados ao pilar Social. Duas empresas (ORG-5 e ORG-8) têm maior similaridade com a palavra 'emissão', sendo que ambas as empresas são do setor mineral e certamente apresentam maior preocupação com o fator Ambiental a respeito da emissão de gases do efeito estufa.

[4] https://www.poder360.com.br/internacional/concentracao-de-gases-do-efeito-estufa-bate-recorde-em-2020-diz-onu/

[5] https://pypi.org/project/gensim/

Figura 12 – Mapa de calor comparando empresas e tópicos (iniciativas) ESG.



Fonte: A autora (2022).

Por outro lado, a ORG-9 foi a única empresa com a palavra 'risco' mais utilizada no relatório indicando provavelmente que o pilar de Governança deve ser o mais representativo no documento. Além disso, os bigramas demonstrados anteriormente para essa empresa estão todos relacionados com o pilar de Governança (Figura 7).

A partir da execução do LDA, foram obtidos os tópicos com as palavras como pode ser visto na Tabela 3. Os rótulos foram atribuídos a partir de uma análise das 15 palavras mais frequentes por tópico. É possível observar uma repetição de palavras por tópicos, como o processo de *tuning* do modelo exige muitas iterações para refinar o modelo e pode ser aperfeiçoado.

Como o LDA se trata de algoritmo de aprendizado não supervisionado de documentos em tópicos, é muito importante ter um breve conhecimento do assunto para determinar o número de tópicos esperados e analisar a coerência das palavras mais relevantes por tópico. Por isso, técnicas complementares como nuve de palavras e N-gramas como utilizados no presente trabalho, podem auxiliar no processo de descoberta desses tópicos.

Tabela 3 – Tabela com os 10 tópicos e palavras mais frequentes.

Tópico	**Palavras-chave**
Topic 1	gri, gesto, so, relatrio, riscos, sobre, governana, aes, sade, trabalho, sustentabilidade, social, segurana, forma, colaboradores
Topic 2	gesto, so, sobre, gri, relatrio, aes, sustentabilidade, trabalho, sade, social, forma, programa, segurana, colaboradores, clientes
Topic 3	relatrio, so, sustentabilidade, gesto, empregados, sobre, aes, social, riscos, gs, impactos, processo, trabalho, segurana, sade
Topic 4	gri, gesto, relatrio, colaboradores, sobre, ita, programa, social, forma, trabalho, so, unibanco, sustentabilidade, sade, informaes
Topic 5	varejo, gri, companhia, relatrio, sobre, gesto, materiais, tica, indicadores, governana, temas, lojas, colaboradores, relacionamento, desempenho
Topic 6	gesto, gri, so, trabalho, colaboradores, relatrio, sobre, aes, programa, segurana, social, sustentabilidade, informaes, sade, governana
Topic 7	gesto, relatrio, sobre, gri, so, colaboradores, aes, sustentabilidade, trabalho, ita, programa, forma, social, anual, segurana
Topic 8	gri, gesto, sobre, relatrio, segurana, riscos, forma, trabalho, aes, social, total, sustentabilidade, so, empresas, colaboradores
Topic 9	gesto, sustentabilidade, sobre, relatrio, anual, social, forma, so, riscos, sade, gri, trabalho, programa, aes, ambiental
Topic 10	gri, gesto, ambiental, social, governana, empresa, material, estratgia, riscos, relato, integrado, segurana, comunidades, contedo, sumrio

4.4 Pós-processamento

Nesta seção apresentam-se os resultados obtidos, dadas as extrações de tópicos e aplicação de métricas. O pacote utilizado para o LDA foi a biblioteca `Gensim` e o `GridSearchCV`[6] que é um módulo do Scikit Learn amplamente usado para automatizar grande parte do processo de *tuning* e identificar o melhor modelo. O objetivo primário do GridSearchCV é a criação de combinações de parâmetros para posteriormente avaliá-las. E posteriormente a biblioteca `pyLDAvis`[7] para visualização.

O parâmetro de ajuste mais importante para modelos LDA é número de tópicos e também o decaimento da taxa de aprendizagem, que é um parâmetro de aprendizado que controla a rapidez com que as informações são esquecidas no modelo (HOFFMAN; BLEI; BACH, 2010).O GridsearchCV constrói vários modelos LDA para todas as combinações possíveis de valores de parâmetro. Portanto, esse processo consome muito tempo e recurso.

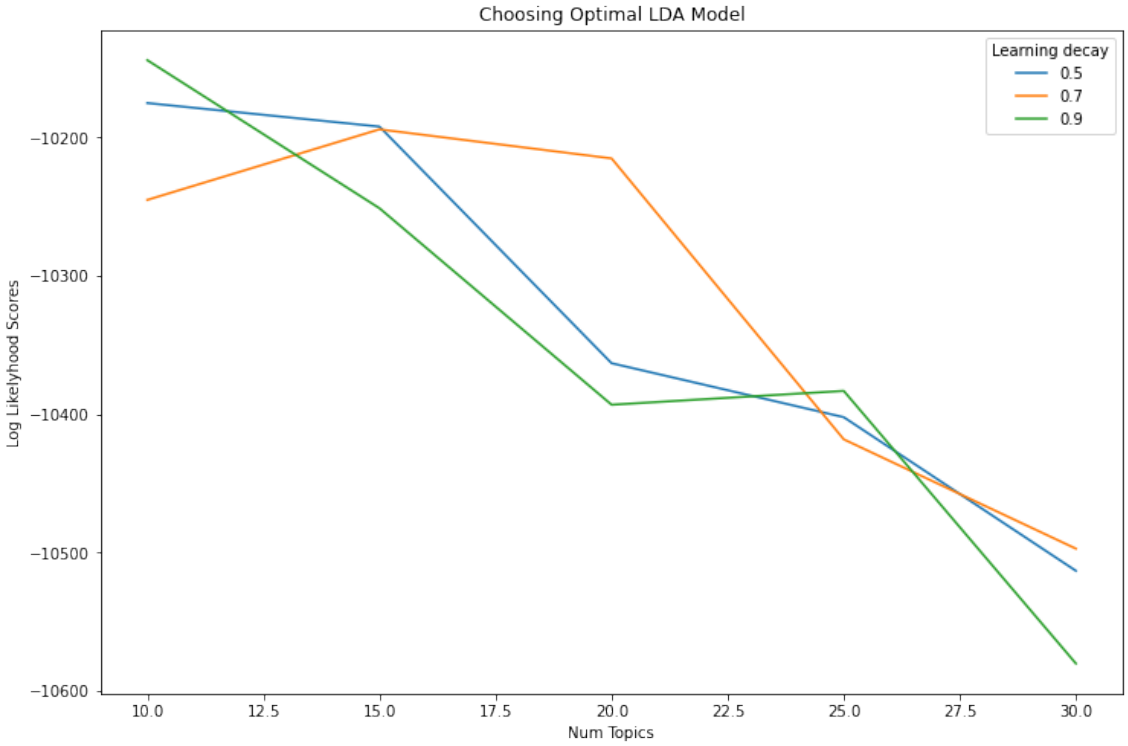
A maneira simples de estimar o modelo probabilístico é encontrar os valores de log-verossimilhança. As pontuações de log-verossimilhança são calculadas para todos os documentos não vistos com a taxa de aprendizado apresentada. O modelo com a pontuação máxima de log-verossimilhança é considerado como melhor modelo. A perplexidade é a maneira de medir de que maneira o modelo é capaz de prever uma amostra. Ele ajuda a determinar um número ideal de tópicos. É calculado tomando o log-verossimilhança de documentos de texto com tópicos resultantes do modelo de tópico. Um modelo satisfatório terá uma alta probabilidade e, consequentemente, baixa perplexidade. Muitas vezes a probabilidade preditiva e opinião humana são menos correlacionadas, portanto adaptável até certo ponto no estabelecimento de negócios.

As combinações testadas foram com 10 a 30 tópicos e o decaimento da taxa de aprendizagem com taxas de 0.5, 0.7 e 0.9. A métrica de perplexidade é bastante utilizada em modelagem de tópicos e quanto menor seu valor melhor o resultado. Os melhores parâmetros encontrados foram o decaimento da taxa de aprendizagem de 0.9 e 10 tópicos e perplexidade igual a 29.02. Os resultados obtidos estão apresentados na figura abaixo (Figura 13).

[6] https://scikit-learn.org/stable/modules/generated/sklearn.model$_s election.GridSearchCV.html$

[7] https://pyldavis.readthedocs.io/en/latest/readme.html

Figura 13 – Gráfico das pontuações de log-verossimilhança em relação ao número de tópicos.



Fonte: A autora (2022).

5 CONCLUSÃO

É notável, que cada vez mais, governos, instituições financeiras e bolsas de valores estão focando nas metas de gestão ESG estabelecidas e na performance das organizações em relação à transparência e à exatidão dos indicadores e relatórios de sustentabilidade. Apesar desta transformação digital possibilitar uma crescente captura de dados, ainda são poucas as empresas que conseguem transformá-los em valor para o negócio fazendo bom uso de tecnologia e análises de dados disponíveis, especialmente quando se trata de dados não-estruturados como em textos.

O presente trabalho trouxe à tona o uso de mineração de textos e modelagem de tópicos para priorizar a classificação das principais iniciativas ESG adotadas por empresas brasileiras para facilitar o reconhecimento das mesmas para implementar uma estratégia que ajude a definir os objetivos, mensurar o desempenho, gerenciar os impactos e identificar os riscos sustentáveis de uma organização. Pois relatórios de sustentabilidade demonstram os impactos positivos e negativos de uma empresa no meio ambiente, na sociedade e na economia. O acompanhamento dos relatórios permite, gradualmente, que a organização encontre oportunidades competitivas e identifique valor para o mercado e, também, para construir confiança e credibilidade por meio de *insights* para gerenciar riscos e oportunidades relacionadas ao fatores ESG.

Afinal, um problema comumente anunciado é a falta de construção da matriz de materialidade das empresas, isso significa, o entendimento dos temas mais importantes para a empresa para definir seu planejamento estratégico. Por exemplo, apesar da relevância de todos os pilares para uma sociedade, instituições financeiras podem ter maior contribuição para os pilares de Governança e Social enquanto empresas do setor industrial tendem para os fatores Ambientais e Sociais. Portanto, isso significa que não necessariamente uma empresa precisa contribuir igualmente para todos os fatores ESG, mas para os pilares mais importantes é necessário reconhecer as oportunidades. Dessa forma, a aplicação de mineração de textos e Processamento de Linguagem Natural pode contribuir cada vez mais para a geração desses *insights*.

Os resultados apresentados são satisfatórios, pois através de uma metodologia clássica de análise exploratória por mineração de texto e modelagem de tópicos através de um algoritmo de Processamento de Linguagem Natural, como o LDA, pode-se observar que os resultados são condizentes com a expectativa, mas também podem revelar alguns aspectos relevantes porém pouco evidentes. Uma vantagem ao utilizar esse tipo de abordagem é a compreensão das tarefas executadas e a análise dos resultados sob a perspectiva humana para avaliar se há coerência nos resultados obtidos.

Contudo, a maior dificuldade encontrada foi o intenso processo de repetições para melhorar os resultados até encontrar o número ideal de tópicos com palavras relacionadas a cada tópico. Este trabalho apresentou as principais palavras por documento através de nuvem de palavras, os bigramas que também representam sequencia de palavras comumente utilizadas nos textos e por fim o resultado do LDA, que demonstrou como melhor modelo 10 tópicos encontrados.

Com a proposta do presente trabalho é possível extrair diversas e valiosas informações dos relatórios de sustentabilidade e ESG divulgados pelas empresas. Contudo, ainda há muitas oportunidades para explorar essas informações e complementá-las com a percepção do mercado. Devido à limitação de tempo e recurso, como trabalho futuro é necessário refinar o modelo apresentado e testá-lo para uma nova base de dados.

Outra possibilidade até para comparação de resultados para modelagem de tópicos é o uso do modelo pré-treinado de redes neurais como o `BERTopic`[1]. Esta técnica de modelagem não foi escolhida por automatizar o processo e ter pouco controle dos parâmetros utilizados além disso, o BERTopic pode performar melhor em textos específicos e a própria limitação de vocabulário dependendo do idioma.

Com um modelo mais aprimorado, uma oportunidade seria a combinação de extração de iniciativas a partir do modelo proposto com a análise de sentimento de notícias ou redes sociais para favorecer uma solução integrada para gestão de informações das empresas na construção da matriz de materialidade.

[1] https://maartengr.github.io/BERTopic/index.html

REFERÊNCIAS

AMEND, A. A data-driven approach to environmental, social and governance. **Engineering Blog**, 2020.

ATKINS, A.; NIRANJAN, M.; GERDING, E. Financial news predicts stock market volatility better than close price. **The Journal of Finance and Data Science**, v. 4, p. 120–137, 2018.

BELINKY, A. Seu esg É sustentÁvel? **GVEXECUTIVO**, v. 20, n. 4, p. 37–44, 2021.

BELLSTAM, G.; BHAGAT, S.; COOKSON, A. Innovation in mature firms: A text-based analysis. **Working paper, University of Colorado Leeds School of Business, Boulder**, p. 1–78, 2017.

BLEI, D.; NG, A. Y.; JORDAN, M. Latent dirichlet allocation. **Journal of Machine Learning Research**, v. 3, p. 993–1022, 2003.

FALEIROS, T. Propagação em grafos bipartidos para extração de tópicos em fluxo de documentos textuais. **Tese**, p. 1–162, 2016.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n. 11, p. 27–34, 1996.

FISHER, I.; GARNSEY, M.; HUGHES, M. A data-driven approach to environmental, social and governance. **Intelligent Systems In Accounting, Finance And Management**, v. 23, n. 3, p. 157–214, 2016.

GUO, T. et al. Esg2risk: A deep learning framework from esg news to stock volatility prediction. **ERN: Stock Market Risk (Topic)**, p. 1–10, 2020.

HAGENAU, M. M.; LIEBMANN, M.; NEUMANN, D. Automated news reading: Stock price prediction based on financial news using context-capturing features. **Decision Support Systems**, v. 55, p. 685–697, 2013.

HOFFMAN, M.; BLEI, D.; BACH, F. Online learning for latent dirichlet allocation. **NIPS**, p. 856–864, 2010.

KIESEL, F.; LüCKE, F. Esg in credit ratings and the impact on financial markets. **Financial Markets, Inst. Inst.**, v. 28, p. 263–290, 2019.

LEWIS, C.; YOUNG, S. Fad or future? automated analysis of financial text and its implications for corporate reporting. **Accounting and Business Research**, v. 49, n. 5, p. 587–615, 2019.

LIDDY, E. **Natural Language Processing**: Surface. Syracuse: Syracuse University, 2001.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **An Introduction to Information Retrieval**. Cambridge: Cambridge University Press, 2009.

MISHRA, S. Esg matters. **Harvard Law School Forum on Corporate Governance**, p. 1–14, 2020.

NUGENT, T.; STELEA, N.; LEIDNER, J. Detecting esg topics using domain-specific language models and data augmentation approaches. **arXiv.org Computer Science**, p. 1–11, 2020.

RAMAN, N.; BANG, G.; NOUBAKHSH, A. Mapping esg trends by distant supervision of neural language models. **Mach. Learn. Knowl. Extr.**, v. 2, n. 4, p. 453–468, 2020.

SERAFEIM, G. Public sentiment and the price of corporate sustainability. **Financial Analysts Journal**, v. 76, n. 2, p. 26–46, 2021.

SOKOLOV, A. et al. Building machine learning systems for automated esg scoring. **The Journal of Impact and ESG Investing**, v. 1, n. 3, p. 39–50, 2021.

UNGARETTI, M. Esg de a a z. **Expert XP**, p. 1–14, 2020.

VAJJALA, S. et al. **Practical Natural Language Processing**. Sebastopol: O'Reilly Media, Inc., 2020.