



**UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS SOCIAIS APLICADAS
DEPARTAMENTO DE RELAÇÕES INTERNACIONAIS
CURSO DE RELAÇÕES INTERNACIONAIS**

IOHAN LUCAS VALE DE MENDONÇA FAUSTINO

A difusão de notícias sobre o encontro do Fórum Econômico Mundial em 2022:
uma abordagem exploratória com Machine Learning e Deep learning.

JOÃO PESSOA

2022

IOHAN LUCAS VALE DE MENDONÇA FAUSTINO

A difusão de notícias sobre o encontro do Fórum Econômico Mundial em 2022:
uma abordagem exploratória com Machine Learning e Deep learning.

Trabalho de Conclusão de Curso apresentado
como requisito parcial à obtenção do título de
Bacharel em Relações Internacionais pela
Universidade Federal da Paraíba.

Orientador (a): Dr. Pascoal Teófilo Carvalho
Gonçalves

JOÃO PESSOA

2022

Catálogo na publicação
Seção de Catalogação e Classificação

F268d Faustino, Iohan Lucas Vale de Mendonca.

A difusão de notícias sobre o encontro do Fórum Econômico Mundial em 2022: uma abordagem exploratória com Machine Learning e Deep learning / Iohan Lucas Vale de Mendonca Faustino. - João Pessoa, 2022.
60 f.

Orientação: Pascoal Teófilo Carvalho Gonçalves.
TCC (Graduação) - UFPB/CCSA.

1. Fórum Econômico Mundial. 2. Aprendizagem de máquina. 3. Modelagem de tópicos. 4. Análise de emoções. I. Gonçalves, Pascoal Teófilo Carvalho. II. Título.

UFPB/CCSA

CDU 327(02)

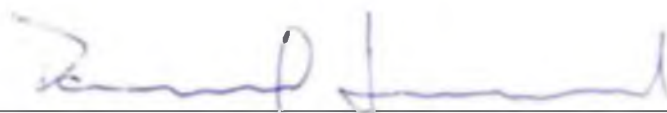
IOHAN LUCAS VALE DE MENDONÇA FAUSTINO

**A DIFUSÃO DE NOTÍCIAS SOBRE O ENCONTRO DO FÓRUM ECONÔMICO
MUNDIAL EM 2022: UMA ABORDAGEM COM MACHINE LEARNING E DEEP
LEARNING**

Trabalho de Conclusão de Curso apresentado
ao Curso de Relações Internacionais do Centro
de Ciências Sociais Aplicadas (CCSA) da
Universidade Federal da Paraíba (UFPB),
como requisito parcial para obtenção do grau
de bacharel (a) em Relações Internacionais.

Aprovado(a) em, 05 de dezembro de 2022

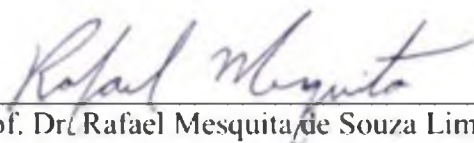
BANCA EXAMINADORA



Prof. Dr. Pascoal Teófilo Carvalho Gonçalves – (Orientador)
Universidade Federal da Paraíba - UFPB



Prof. Dr. Túlio Sérgio Henriques Ferreira
Universidade Federal da Paraíba - UFPB



Prof. Dr. Rafael Mesquita de Souza Lima
Universidade Federal de Pernambuco - UFPE

Agradecimentos

Gostaria primeiramente de agradecer ao orientador Professor Dr. Pascoal Gonçalves por apoiar o projeto de monografia. Em seguida, gostaria de agradecer ao professor Professor Dr. Lucas Milanez, com quem fui pesquisador em dois projetos de Iniciação Científica, e o professor Dr. Iure Paiva, com quem fui pesquisador em uma iniciação tecnológica. Ambos professores foram de fundamental importância na construção do aprendizado científico e do exercício crítico que baseia uma análise científica.

Além dos professores, gostaria de agradecer ao professor Dr. Guerd Rocha e aos doutorandos Igor Grillo e Iran Silva do Departamento de Química da UFPB pelo suporte dado para realização computacional da pesquisa. Somado a eles, gostaria de agradecer ao Centro Nacional de Processamento de Alto Desempenho do Estado de São Paulo pela disponibilização de máquinas para realização computacional da pesquisa.

Gostaria de agradecer a todos meus amigos, colegas, demais professores do Departamento de Relações Internacionais e Economia e familiares que me acompanharam nessa trajetória acadêmica. Em especial, gostaria de agradecer aos professores Dr. Wagner Faustino e Dra. Juliana Vale, meus pais, e Dra. Haryane Moraes e Me. José Filho, minha madrasta e meu padrasto. Esses últimos quatro nomes foram altivos e ativos na minha formação como pessoa e pesquisador.

*"We feel free because we lack the very language to articulate our unfreedom."
(Slavoj Žižek)*

Resumo

O Fórum Econômico Mundial é uma organização internacional voltada para cooperação internacional público-privadas. Desde 1970, a instituição passou por diversas mudanças que a levou a ter reconhecimento e prestígio internacional entre as elites internacionais. Segundo Khanna, o Fórum é um empreendedor de normas, facilitador e convocador em uma estrutura baseada nas partes interessadas. O fórum recebe por meio de seu encontro anual em Davos (Suíça) grande atenção mundial ao reunir lideranças de organizações públicas de organizações internacionais, representantes corporativos privados, ativistas, lideranças da juventude, cientistas e a mídia. Por consequência, um grande volume de informações é produzido e disponibilizado na internet. Tendo como base esses dados, a seguinte pesquisa tem por objetivo geral explorar a difusão global de notícias em termos de tópicos e sentimentos predominantes. Toda metodologia utilizada é baseada em Machine Learning e Deep Learning, tendo como foco modelos da família LDA, para modelagem de tópicos, e BERT para análise de emoções. Os resultados sugerem que o Corpus pode ser dividido em 9 tópicos, e que o interesse primário, secundário e terciário pode variar de acordo com os países. Além disso, observou-se que o sentimento predominante foi de "Alegria", seguido de "Raiva". Ao fim, também pôde se constatar que mais da metade dos Corpus de cada país pode ser explicado em até três tópicos predominantes.

Palavras-chave: Fórum Econômico Mundial; Aprendizagem de Máquina; Modelagem de Tópicos; Análise de Emoções.

Abstract

The World Economic Forum is an international organization focused on international public-private cooperation. Since 1970, the institution has undergone several changes that have led it to have international recognition and prestige among international elites. According to Khanna, the Forum is a norm-setter, facilitator, and convener in a stakeholder-based structure. Through its annual meeting in Davos (Switzerland) the Forum receives great worldwide attention as it brings together leaders of public organizations of international organizations, private corporate representatives, activists, youth leaders, scientists, and the media. Consequently, a great deal of information is produced and made available on the Internet. Based on this data, the following research has the overall goal of exploring global news dissemination in terms of topics and prevailing sentiments. All methodology used is based on Machine Learning and Deep Learning, focusing on models from the LDA family, for topic modeling, and BERT for emotion analysis. The results suggest that the Corpus can be divided into 9 topics, and that primary, secondary and third interest can vary across countries. Furthermore, it was observed that the predominant feeling was "Joy", followed by "Anger". In the end, it could also be found that more than half of the Corpus from each country can be explained in up to three predominant topics.

Keywords: World Economic Forum; Machine Learning; Topic Modeling; Emotion Detection.

Lista de ilustrações

Figura 1 – Gráfico de notícias coletadas por geolocalização	25
Figura 2 – Figura do plate notation do LDA	28
Figura 3 – Figura de Redes Neurais Artificiais	30
Figura 4 – Sliding Window	31
Figura 5 – Gráfico comparativo entre os modelos ProdLDA, NeuraLDA e LDA .	37
Figura 6 – Gráfico do RTPC para LDA	38
Figura 7 – Distribuição de probabilidade condicional de palavras por tópico (modelo 5 tópicos)	39
Figura 8 – Distribuição de probabilidade condicional de palavras por tópico (modelo 9 tópicos)	40
Figura 9 – Distribuição de probabilidade condicional de palavras por tópico (modelo 12 tópicos)	41
Figura 10 – Distribuição de frequência das 10 maiores palavras por tópico entre os 20 documentos mais representativos	43
Figura 11 – Distribuição absoluta de documentos por tópicos	45
Figura 12 – Distribuição relativa de tópicos por países	46
Figura 13 – Distribuição relativa de emoções por países	48
Figura 14 – Distribuição relativa de emoções por tópicos	49

Lista de tabelas

Tabela 1 – Tabela dos hiperparâmetros para LDA	33
Tabela 2 – Tabela dos hiperparâmetros para ProdLDA e NeuraLDA	33
Tabela 3 – Síntese do modelo 5	38
Tabela 4 – Síntese do modelo 9	44

Lista de abreviaturas e siglas

FEM	Fórum Econômico Mundial
OI	Organização Internacional
UFPB	Universidade Federal da Paraíba
LDA	Latent Dirichlet Allocation
RNA	Redes Neurais Artificiais
BERT	Bidirectional Transformers for Language Understanding

Lista de símbolos

α	Hiperparâmetro de uma Distribuição Dirichlet
β	Hiperparâmetro de uma Distribuição Dirichlet
θ	Parâmetros de uma distribuição de palavras por tópicos
π	Parâmetros de uma distribuição de palavras por tópicos
w	Palavra
V	Vocabulário
i	Conjunto de palavras em um documento
d	Documento original
$\hat{d}$	Documento gerado
C	Parâmetro de uma distribuição multinomial de associação de palavra por tópico

Sumário

	Lista de ilustrações	5
	Lista de tabelas	6
1	INTRODUÇÃO	10
2	FORUM ECONÔMICO MUNDIAL E MÍDIA: UMA ABORDAGEM MULTIDISCIPLINAR	12
2.1	Uma breve retrospectiva histórica do Fórum Econômico Mundial	12
2.2	Características do Forum: regime, funções, agentes e a mídia	16
2.3	Mídia: como investigar sua atuação ?	19
3	METODOLOGIA	24
3.1	Revisão de literatura	24
3.2	Obtenção de dados e pre-processamento	24
3.3	Modelagem de tópicos	26
3.3.1	Os modelos selecionados: LDA, ProdLDA e NeuraLDA	26
3.3.2	Problema da otimização, métrica de qualidade e nomenclatura de tópicos	31
3.4	Análise de emoções por redes neurais	34
4	EXPLORAÇÃO ANALÍTICA DA PESQUISA	36
4.1	A seleção entre tipos de modelos de tópicos diferentes e a escolha do número de tópicos	36
4.2	Resultados da modelagem de tópicos	45
4.3	Resultados da modelagem de emoções	47
5	CONCLUSÃO	50
	REFERÊNCIAS	53

1 Introdução

O Fórum Econômico Mundial (FEM) é a organização internacional voltada para a parceria público-privada. Iniciado na década de 70, essa instituição passou por profundas transformações elevando sua capacidade de ação e engajamento, produzindo um nicho diverso de atores. Nesse sentido, Khanna (2010) acredita que o Fórum serve como uma plataforma para catalisar ação, com funções de convocador, facilitador e empreendedor de normas. Conjuntamente com outras Organizações Internacionais (OIs), essa instituição procura construir a agenda global de cooperação, sob a égide da cooperação público-privada.

Anualmente, o FEM promove seu mais importante encontro: o evento de Davos (Suíça). Nesse encontro, participam as elites internacionais, representantes do terceiro setor, governos e outras OIs. Conjuntamente, esses atores atraem muita atenção da mídia tradicional de diversos países, sendo ela também um stakeholder do FEM. Por conseguinte, muitas notícias são compartilhadas em diferentes países.

Produzindo um grande volume de dados textuais por via de plataformas online sobre os acontecimentos do evento, a mídia tradicional atua como um difusor de informações, como um mecanismo capaz de produzir prestígio e como um membro que engaje a opinião pública sobre problemas internacionais (PIGMANN, 2007; KHANNA, 2010; FRIESEN, 2017).

Baseado nisso, o presente estudo tem como objetivo geral explorar a difusão de notícias entre uma seleção de países membros em termos de tópicos e quais sentimentos predominantes foram evocados. Para alça-lo, perseguem-se os objetivos específicos de contextualizar o Fórum Econômico Mundial e o papel da mídia, explorar a difusão de notícias por tópicos através de técnicas de Machine Learning e Deep Learning e explorar as emoções predominante evocadas através de modelo de Deep Learning.

Para operacionalizar esses objetivos este trabalho se estende em mais 4 capítulos posteriores. No capítulo 2 é abordada a problemática do Fórum Econômico Mundial em perspectiva histórica, a sua conceituação em Relações Internacionais e os estudos multidisciplinares para entender o papel de contextualizar a abordagem utilizada na investigação da mídia.

No capítulo 3 está a metodologia utilizada nesse trabalho. Nela se discute como os dados foram obtidos e tratados antes da análise, dos modelos utilizados para modelagem de tópico e detecção de emoções e da revisão bibliográfica que sustentou a fundamentação. Por conseguinte, o capítulo 4 aborda a problemática da seleção do

modelo e interpretação dos resultados obtidos para tópicos e detecção de emoções.

Apois mais de 50 mil simulações comparando-se três tipos de modelos de modelagem de tópicos e os demais procedimentos explicitados na metodologia, obteve-se um modelo que pôde representar o corpus em 9 tópicos. Esse modelo foi útil para descobrir que o interesse primário, secundário e terciário muda de acordo com os países, sendo esses países não necessariamente da mesma região. Pôde também se observar que mais de 50% da cobertura de todos os países pode ser explicada em três tópicos. Somado a isso, ao jogar luz à análise de emoções, pôde se observar que "Alegria" foi o sentimento geral do evento, menos quando o assunto estava relacionado à fome ou guerra da Rússia e da Ucrânia. Demais observações podem ser encontradas no capítulo de análise exploratória da pesquisa.

Finalmente, o capítulo 5 traz uma síntese dos principais pontos, das limitações e de possíveis desdobramentos futuros.

2 Fórum Econômico Mundial e Mídia: uma abordagem multidisciplinar

Este capítulo tem por objetivo apresentar um panorama geral sobre (2.1) a descrição do desenvolvimento do Fórum Econômico Mundial (FEM) como organização internacional para parcerias público-privadas em perspectiva histórica (2.2) a conceituação do Fórum e do papel da mídia na literatura específica em Relações Internacionais e (2.3) diferenciar os estudos clássicos e apresentar a base multidisciplinar que sustenta o tipo de investigação sugerida. Todos esses procedimentos são partes do embasamento contextual do objeto e da forma de análise do agregado dos agentes sobre um tema específico: o encontro de Davos em 2022.

2.1 Uma breve retrospectiva histórica do Fórum Econômico Mundial

O Fórum Econômico mundial existe desde 1971 como uma organização sem fins lucrativos que reúne celebridades, lideranças públicas e privadas de todo o mundo para discutir problemas que tangem política, negócios e cultura. Nos três primeiros anos, foram discutidos como as empresas europeias poderiam realizar o catch-up conjuntamente aos Estados Unidos, mais especificamente como através da abordagem dos stakeholders (PIGMAN, 2007, p.9).

Essa abordagem, que se seguirá ao longo da história da instituição, foi diretamente influenciada pelas contribuições teóricas de Klaus Schwab's. Sobre a história da instituição, Pigman enfatiza que

Para serem eficazes na maximização do potencial de uma empresa, os gerentes precisam levar em conta os interesses de todos os stakeholders da empresa: não apenas acionistas, mas clientes finais e intermediários, funcionários, equipe gerencial e os interesses mais amplos das comunidades nas quais a empresa está situada, incluindo vizinhos na proximidade imediata da empresa, governos e outros usuários do ambiente em que a empresa opera. [...] Schwab imaginou que os gerentes das empresas europeias, ao adotarem a abordagem das partes interessadas, poderiam aumentar seu jogo de forma competitiva em resposta ao desafio colocado por seus desafiantes americanos mais eficientes.(PIGMAN, 2007, p.9, tradução nossa)

Mantendo constante essa abordagem, a instituição passou por uma série de mudanças significativas. Sobre o engajamento na instituição, Pigman (2007) enfatiza

que ambos boicotes das produções dos países membros da OPEC e colapso do sistema de Bretton Woods foram fatores que engajaram a participação no FEM. Nesse contexto, o Fórum foi importante por aproximar líderes empresariais do oriente médio e do ocidente, tendo impacto direto nas negociações entre comunidade europeia e o Conselho de Cooperação do Golfo . Esse período é também importante pela adesão de países em desenvolvimento e pela adesão da República Popular da China, em 1979. Nesses primeiros momentos, a instituição já se mostrou eficaz em aproximar parceiros públicos e privados, incluindo atores fora do espectro ocidental.

A década de 80 é marcada pelo crescimento da instituição e a adesão global. Dentre os acontecimentos mais significativos da década está a aproximação do General Agreement on Tariffs and Trade (GATT) e do Fórum Econômico Internacional (FMI) ao FEM. Fruto da aproximação está o papel central do encontro de Lousanne (Suíça) em 1982, em que ministros representantes do GATT e empresários deram marcha à rodada do Uruguay em 1986 (PIGMANN, 2007, p.13). Esse período também é importante pela entrada da Índia em 1985. Outro marco é a mudança do nome do Fórum, que deixou de ser European Management Forum para ser World Economic Forum. Esse nome exprime também a concepção de atuação do Fórum, visando solucionar problemas que iriam além da Europa.

A partir da década de 90, Davos já havia se solidificado à nível internacional, tendo sua agenda moldada pelos acontecimentos presentes do período. Em especial, cabe destacar o desmoronamento da União Soviética, o avanço da integração financeira internacional e suas instabilidades em relação à crise das dívidas e a inclusão cultural de novos nomes da elite global (MARSHALL, 2015). Esses acontecimentos repercutiram nos eventos, onde discutiu-se temas como necessidade de adoção de medidas austeras visando a criação de um sistema financeiro internacional mais confiável e à abertura de novos mercados e oportunidades na África, Europa Oriental e Ásia (MARSHALL, 2015). Até 1999, a elite global começava a se integrar em torno de Davos e discutir sobre o futuro da cooperação público-privado e da globalização para o século seguinte.

Entre 2000 e 2010, Davos vivenciava um momento de muita atenção das autoridades políticas. Em 2000, o presidente Clinton inaugura a presença americana no evento (PARKER, 2019). Em 2007 a Rússia demonstra aproximação indicando o primeiro-ministro Dmitry Medvedev até que em 2009 o presidente Putin vai pela primeira vez. Em 2009, Putin salientava o papel crucial da Rússia na garantia da segurança energética global; os interesses nacionais para a sociedade civil; e respondia que seguiria com a política de “Campeões Nacionais”, como outros países vinham realizando (DAVOS, 2009).

Em 2005, Lula discursa em Davos e defende o combate à fome; cita a

importância do Brasil na constituição do World Social Fórum, instituição crítica à WEF; e apresenta ao mundo o compromisso com reformas tributária e da previdência e outras leis como a de Parceria Público-privada e de Falências (DAVOS, 2005). Davos agora conseguia aproximar a elite nacional e internacional com os líderes estatais, onde ambas as partes poderiam expor sua visão, debater, discordar, tirar dúvidas.

Nessa década, problemas de naturezas bastante diversas foram discutidos no fórum. Em 2002, o evento foi celebrado em Nova York em homenagem ao 11 de setembro. Esse evento lançou luz sobre o potencial dos movimentos terroristas internacionais, mas ao mesmo tempo também discutia o combate a doenças contagiosas, como HIV (WORLD ECONOMIC FORUM, 2002). Temas como negócios e finanças, desenvolvimento econômico e crescimento, saúde e geopolítica e governança global se mantiveram em perspectiva ao longo da década como apontam os próprios relatórios anuais de 2002 e 2008.

Além disso, o Fórum começou a implementar ações voltadas à responsabilidade social ao lançar, em 2001, o Relatório de Governança Global. Esse relatório pretendia “classificar os esforços de governos, entidades intergovernamentais e atores não estatais para abordar problemas globais que vão da pobreza à doença, degradação ambiental ao conflito armado” (FRISEN, 2020, p.73, tradução nossa). O Fórum, nesse sentido, iniciava um processo de normatização de seus participantes para atingir objetivos reais em termos de agenda global (KHANNA, 2010)

Ainda entre 2000 e 2010, o FEM fortalece a criação de redes entre os stakeholders. Parte desse feito se deve a criação da rede de Conselhos da Agenda Global (CAG), em que cada um dos 68 GACs era liderado por um conjunto de experts buscando uma base comum de gerenciamento de risco de negócios para problemas pontuais da agenda global (FRISEN, 2020).

Em 2009, o Fórum lança a iniciativa da criação dos Relatórios de Iniciativas de Redesenho Global (RIRG), que consistia em 58 propostas para fortalecer a cooperação internacional (FRISEN, 2020). Em consonância, no mesmo foi lançado o Parceria de Agenda Global (PAG), que pretendia ser uma plataforma conjunta entre empresas privadas e governos para avaliar riscos sistemáticos e vulnerabilidades de estratégias e políticas públicas (FRISEN, 2020). Essas três iniciativas do final do período demonstram o desejo de Davos de fortalecer e operacionalizar o papel institucional na formulação da agenda global.

Dotado de participação pública em diferentes âmbitos e privada de diferentes naturezas e setores, o Fórum conseguiu engajar públicos em torno de problemas específicos para cada setor interessado. A segmentação de espaços de interesse combinada com a união em torno de princípios comuns moldou o comportamento da instituição entre 2010 e 2020. Em especial, cabe destacar o encontro de 2016,

com o lançamento das iniciativas de sistemas. Essa ideia é definida pelo criador da instituição como

a ideia de que mudanças positivas sustentáveis pode ser alcançado através de uma compreensão contextual de como o componente partes de um sistema complexo interagem e influenciam o sistema nos pontos mais efetivos – está agora orientando como o Fórum emprega seus recursos limitados para ter um impacto global (WORLD ECONOMIC FORUM, 2016, p.3, tradução nossa).

Paralela à abordagem de engajamento para mudanças, o Fórum foi palco para discursos de política internacional. Em 2017, é notável lembrar do discurso enfático do presidente Xi Jinping em defesa do livre comércio e contra a guerra comercial, levantando uma crítica aos Estados Unidos por estarem com um comportamento protecionista que prejudicaria o potencial da globalização (BARKIN e PIPER, 2017). Em 2019, o Secretário Geral da ONU, Antônio Guterres, apontou a combinação da globalização entre avanços para o mundo, e ainda carregar sérios problemas que se repetiam do passado, como a crise dos débitos e a desigualdade interna dos Estados (GUTERRES, 2019). Guterres (2019) também enfatiza para o crescente papel das Notícias Falsas e da desinformação como processo que contribui para o descrédito das instituições. Nesse mesmo ano, o Fórum (2019b) destaca o decréscimo da credibilidade da mídia tradicional, agente ativamente participativo dos encontros em Davos.

Se antes o clima era engajamento e transformação, em 2021 o tema abordado foi “O grande recomeço”, partindo do diagnóstico de que a

COVID-19 demonstrou que nenhuma instituição ou indivíduo sozinho pode enfrentar os desafios econômicos, ambientais, sociais e tecnológicos de nosso mundo complexo e interdependente. A pandemia acelerou mudanças sistêmicas que eram aparentes antes de seu início. As falhas que surgiram em 2020 agora aparecem como uma encruzilhada crítica em 2021. O momento de reconstruir a confiança e fazer escolhas cruciais está se aproximando rapidamente, à medida que a necessidade de redefinir as prioridades e a urgência de reformar os sistemas se fortalecem em todo o mundo. (WORLD ECONOMIC FORUM, 2021, tradução nossa)

Em especial, nesse ano houve uma preocupação com as chamadas cadeias de valor. Nesse âmbito, três pontos foram discutidos: a fragilidade das cadeias de suprimentos, colaboração para o suprimento global de alimentos e tecnologia e startups voltadas para as necessidades das cadeias de suprimentos (MONK, 2021). Além disso, houve uma preocupação no evento sobre saúde mental e trabalho; combate à pandemia da Covid-19; uso de tecnologias para a redução de emissão de CO₂, necessidade do reflorestamento, limpezas do oceano e a construção de cidades

sustentáveis; e a economia circular para desenvolvimento nos países de renda-média (WORLD ECONOMIC FORUM, 2021).

Apesar da seriedade do encontro, todavia, o tema de 2021 foi marcado por uma série de teorias conspiratórias e notícias falsas, como o movimento antivacina e a teoria da “nova ordem mundial” (ROBINSON et al, 2021). De fato, experimentou-se nesse evento algo que Guterres já havia avisado em 2019: o crescente movimento antiglobalização, influenciado em parte pelo poder das notícias falsas (GUTERRES, 2019).

Em perspectiva histórica, é possível notar que o FEM mudou consideravelmente desde sua criação. Não apenas o nome da instituição foi transformado, mas a quantidade e natureza de seus membros, sua forma de atuação e sua agenda. A instituição foi importante por aproximar lideranças globais, elites empresariais, representantes das sociedades civil, lideranças da juventude, de Estado e de organizações internacionais buscando solucionar problemas comuns através de ação coletiva. Nesse aspecto, a instituição, apesar de natureza distinta de outras instituições como a ONU, é bem sucedida como um fórum em busca de uma coesão entre parceiros para solucionar problemas globais.

2.2 Características do Forum: regime, funções, agentes e a mídia

Na literatura de Relações Internacionais, Krasner sustenta que regimes internacionais são o conjunto de “princípios, normas, regras e procedimentos decisórios implícitos ou explícitos, em torno dos quais convergem as expectativas dos atores em uma determinada área das relações internacionais” (KRASNER, 1983, p.2, Tradução Nossa). À guisa dessa definição, é possível compreender regimes internacionais privados como um subtipo de regimes internacionais, mas com abordagens distintas. “Esses regimes do setor privado abordam os problemas de eficiência internacional, a segurança ou estabilidade dos mercados, o poder e a autonomia das empresas e a inserção social dos atores econômicos” (HIGGOTT et al, 2000, p.122). Fica evidente nesses regimes que são os atores não-estatais que cooperam e pautam uma agenda própria.

Ainda que com abordagens distintas, ambos membros de regimes público e privados podem cooperar e construir uma governança através das Parcerias Público-Privadas (PPPs). Por PPPs compreende-se “a formação de relações cooperativas entre o governo, empresas com fins lucrativos e organizações privadas sem fins lucrativos para cumprir uma função política” (LINDER e ROSENAU, 2000, p.5, tradução nossa). É nesses espaços cooperativos acontecem as “interações

transnacionais”, compreendidas como “o movimento de itens tangíveis ou intangíveis através das fronteiras do estado quando pelo menos um ator não é um agente de um governo ou uma organização intergovernamental” (NYE e KEOHANE, 2012, p.332). Essas interações transnacionais no contexto de Davos envolvem desde alocação de recursos para financiamentos de projetos à troca de informações sobre assuntos comuns.

Parte dos atores são considerados *sovereignty-free actors* e constituem universo multicentrado de atores autônomos que buscam alcançar parcialmente objetivos de interesse internacional (ROSENAU, 1999, p.36). Esses agentes são elementos de construção dos próprios regimes privados internacionais capazes de criar suas esferas de autoridade autolegitimadas e autônomas (KHANNA, 2010, p.19). Nesse contexto, a interação entre eles pode ser definida em ambos termos de autoridade e barganha. Por autoridade compreende-se o envolvimento seletivo e hierárquico de agentes capazes de exercer sua autoridade com aceitação oriunda de sua legitimidade (ROSENAU, 1999, p.192), como é o caso da seleção dos parceiros estratégicos na construção da agenda da instituição do FEM. Em termos de barganha, a interação entre os agentes é dada em termos de ofertas e contra-ofertas em um espaço delimitado (ROSENAU, 1999, p.192, tradução nossa), como é o caso dos grupos industriais e das deliberações das comunidades epistêmicas do próprio Fórum.

Dadas suas características intrínsecas de seus atores e extrínsecas no âmbito dos regimes, Khanna (2010) afirma que o FEM pode ser visto como um Fórum para realização de parcerias público-privadas com papel de convocador, facilitar e empreendedor de normas. Nesse contexto, sobre o papel convocador, Davos permite

uma oportunidade para examinar até que ponto a participação do FEM se ampliou para além da esfera corporativa, se o FEM promoveu com sucesso uma comunidade epistêmica de múltiplos atores e líderes, se foi capaz de moldar o conteúdo e a direção da “agenda global” entre esses líderes e se influenciou os resultados políticos e institucionais internacionais oficiais.” (KHANNA, 2010, P.64)

A construção de uma comunidade epistêmica passa pela ideia de que novos atores foram ganhando legitimidade para poder participar ativamente do diálogo, ampliando o escopo puramente corporativo para autoridades governamentais e de organizações internacionais, lideranças da juventude, acadêmicos e ativista, como visto na seção anterior. À medida que entravam novos membros e participando ativamente nos encontros anuais, atores foram ganhando legitimidade para participar e coproduzir a própria comunidade epistêmica do fórum (KHANNA, 2010, p.66-69), totalizando 38 comunidades atualmente (WORLD ECONOMIC FORUM, 2022).

Foi justamente o modelo inclusivo que permitiu a criação do próprio espaço de legitimidade dos atores e de suas práticas de legitimação (KHANNA, 2010, p.67). Essas comunidades se dividem em vários espaços temáticos que lhes garantem mais autonomia no Fórum. Assim, o papel de convocador pode ser visto ao criar um ambiente propício ao diálogo entre atores com capacidades reais e simbólicas diferentes em torno de princípios, normas e processos de tomada de decisão comum, mesmo que com menor poder normativo que as organizações de regimes internacionais públicos.

Além de convocador o Fórum tem o papel de facilitador ao condensar capacidade de setores transnacionais e transeitoriais (KHANNA, 2010, p.95-96). A instituição muitas vezes atua como mediadora de interesses e engaja seus membros a tomarem iniciativas conjuntas e políticas regionais, produzindo um input para deliberações intergovernamentais (AVANT et al, 2010). Khanna argumenta que são evidências de seu poder facilitador: a defesa incisiva no potencial bilionário do financiamento da transição energética, dentro das Iniciativas de Mudança Climática; a mediação entre gigantes farmacêuticas e financiamento do fundo global de combate à AIDS, dentro das Iniciativas de Saúde Global; e a difusão de material anticorrupção para empresas, definição de regras para “neutralidade competitiva” industrial e a demanda por apoio da OCDE e ONU contra corrupção, no âmbito das Iniciativas de Parceria Contra Corrupção.

Além de convocador e facilitador, o Fórum pode ser visto como um empreendedor de normas. O maior exemplo desse papel é o da Iniciativa de Redesign Global (IRD) que uniu autoridades nacionais, subnacionais, de organizações internacionais, empresas privadas e representantes da academia com o objetivo de estabelecer normas e gerar ação coletiva para superar a crise de 2008 (KHANNA, 2010, p.162-164). O IRD alcançou o comprometimento direto de Suíça, Catar e Singapura com sua agenda e apoio parcial de lideranças do G20. Ambos atores se comprometeram, ainda que parcialmente, para adotar padrões de comportamento e cooperação para que a crise de 2008 não viesse a se repetir (KHANNA, 2010).

Além dos três papéis supracitados, a própria instituição atua como monitor, produzindo conteúdo gratuito e acessível sobre o cumprimento de compromettimentos do Fórum. Evidências empíricas de seu papel podem ser vistas através de sua abordagem das iniciativas, especificamente a “Iniciativa de Governança Global, para servir como um monitor independente do progresso global em direção ao prazo de 2015 para alcançar os Objetivos de Desenvolvimento do Milênio das Nações Unidas” (PIGMAN, 2007, p.17).

No contexto dos regimes de parceria público-privadas, de atores não-governamentais diversos e dos papéis exercidos pelo FEM, a mídia pode

ser vista como mais um stakeholder da instituição. Inicialmente, esse tipo de stakeholder participava de forma informal, ainda na década de 80 (PIGMAN, p.13), mas logo depois foram chamados para compor o evento de Davos, quando em 2010 cerca de 150 lideranças midiáticas são convocadas, dentre eles jornalistas, âncoras de televisão, repórteres e representantes de blogs (KHANNA, 2010, p.80).

Assim, segundo KHANNA (2010, p.80) a mídia criou seu próprio espaço de legitimidade no evento, servindo não apenas para trazer transparência, mas também para trazer visibilidade tanto para o evento como para outros stakeholders. Nesse sentido Khanna converge com Pigman ao entender que “o papel mais importante que a mídia desempenha em Davos é servir como canal de comunicação para os participantes que desejam se comunicar com o público global” (PIGMAN, 2007, p.78).

2.3 Mídia: como investigar sua atuação ?

Estudos que investigam Cobertura Midiática, Comunicação e Relações Internacionais podem ser bastante diversos em ambos epistemologia, metodologias e objetos de análise. Em termos qualitativos, ressalta-se a sistematização proposta por Gilboa (2002) que leva em consideração categorias para explicar linhas de pesquisa sobre o papel da mídia como elemento de Soft Power na construção de política internacional. Além desse trabalho, ressalta-se no campo quantitativo, baseado em estatística clássica, o trabalho de Wanta, Golan e Lee (2004) sobre a investigação da influência da cobertura midiática sobre outros países na percepção pública americana.

Em “Global communication and foreign policy”, Eytan Gilboa (2002) estabelece sistematicamente como a mídia pode atuar na formulação e condução da política internacional. No trabalho, Eytan traz um insight importante de Nye e Owens (1996, p.21) sobre o papel da comunicação internacional como elemento de “soft power” na construção da política internacional. Ressaltado esse ponto, Gilboa procura entender como os meios de comunicação podem atuar nesse contexto para atingir objetivos desejados em termos de política internacional.

No artigo, Gilboa (2002, p.733) elenca a taxonomia de atores e conceitos, dividindo em 4 categorias (tipo de ator, atividade, contexto e conceito). Na taxonomia proposta, existe o tipo de ator controlador, que supostamente substituem os formuladores de políticas no contexto de intervenções militares-humanitárias, cuja teoria remete à “CNN effect”.

Além desse tipo, existe o tipo de ator que restringe o formulador de política no contexto de tomada de decisão, cujo teoria remete à “real-time policy”; o tipo de ator interventor que se torna mediador internacional em um contexto de solução de conflitos, cujo teoria remete à “international political brokerage”; e o tipo de

ator instrumental que promove negociações e acordos na solução de conflito, cujo teoricamente a “Media Diplomacy”.

Em síntese, Gilboa (2002, p.733-740) sugere que a teoria de “efeito CNN” (“CNN Effect”) deriva de uma reflexão de formuladores de políticas do papel dominante dos grandes conglomerados da mídia, em especial a CNN, na formulação da política internacional após a guerra fria.

Posteriormente, Gilboa sustenta que o ator que restringe (Constraining Actor) acontece devido ao grande volume e velocidade que os meios de comunicação possuem na transmissão de informações, levando à produção de sugestões em política internacional com informações incompletas por parte dos analistas (GILBOA, 2002, p.736-738). Ademais, o autor sugere que o conceito de ator interventor (“Intervening Actor”) está relacionado ao papel ativo que repórteres tiveram na mediação de conflitos. Finalmente, o autor caracteriza os atores instrumentais (Instrumental Actor) como aqueles usados por líderes para atingir fins em política internacional (GILBOA, 2002, p.741-743; GILBOA, 1998, p.62-63).

Amplamente citado, o artigo de Gilboa é fundamental na sistematização da mídia avaliando como ela pode influenciar ou ser usada no processo de formulação de política internacional. Todavia, existem outras linhas de pesquisa que procuram pensar o papel da mídia, da opinião pública e da formulação de política no âmbito das relações internacionais.

Nesse contexto, Wanta, Golan e Lee (2004) em “Agenda setting and international news: Media influence on public perceptions of foreign nations” procuram quantitativamente mensurar o papel da mídia na percepção de nações estrangeiras investigando em especial os Estados Unidos. Os autores partem de duas hipóteses de pesquisa: (H1) Quanto mais cobertura global da mídia uma nação recebe, mais os indivíduos pensarão que é de vital importância para os interesses dos EUA. [...] (H2) Quanto mais cobertura negativa da mídia uma nação recebe, mais os indivíduos pensarão negativamente sobre aquela nação. Quanto mais cobertura positiva da mídia uma nação recebe, mais os indivíduos pensarão positivamente sobre a nação. (WANTA, GOLAN e LEE, 2004, p.369)

Para testar as hipóteses, os autores entrevistaram 1507 pessoas para descobrir a reação delas defronte de notícias de agenda pública do Chicago Council for Foreign Relations, visando instrumentalizar a primeira hipótese; e notícias das empresas ABC, CBS, NBC e CNN para instrumentalizar a segunda hipótese (WANTA; GOLAN; LEE, 2004, p.369-370). No caso da segunda hipótese, os autores codificaram manualmente os sentimentos e testaram compararam com um codificador independente. Como resultado da pesquisa, os autores concluíram, estatisticamente, que existem uma relação entre a cobertura da mídia e como as pessoas veem as nações; e que apenas

as reações negativas tiveram significância estatística provando a relação que a mídia pode influenciar negativamente como as pessoas veem as nações.

Os 2 trabalhos supracitados são mais relacionados a trabalhos clássicos ou quantitativos ou qualitativos. Além desses trabalhos, existem outros trabalhos em outras disciplinas que podem ser considerados pertinentes a relações internacionais na perspectiva da mídia como objeto de análise. Esses estudos selecionados abrangem frameworks diferentes, mas têm em comum a modelagem estrutural de tópicos como vetor comum analítico.

Em estudos em segurança cujo objeto de estudo são organizações terroristas como Kfir (2017), Fainberg (2017) e Vasiliev et al (2018) procuram-se investigar sobre a identidade e a comunicação do Estado Islâmico, considerado como elemento principal de Soft Power (VASILIEV et al, 2018). Nesse contexto, estudos como Ghajar-Khosravi et al (2016) e Maragkos e Maravelakis (2022) trazem insights importantes, baseado em dados de mídias sociais, de como ocorre a comunicação do Estado Islâmico é feita a nível sujeito e do coletivo.

Ghajar-Khosravi et al (2016) buscaram explorar como diferem a comunicação na rede social entre apoiadoras do Estado Islâmico e não apoiadoras, usando uma amostra de membros do Estado Islâmico como controle. Para realizar a pesquisa, coletou-se dados do Twitter de apoiadoras, não apoiadoras e grupo de controle. Como resultado, observou-se no trabalho que as amostras de apoiadoras e não apoiadoras diferiam pelas palavras e tópicos que eram mais frequentes nos posts.

Por sua vez, Maragkos e Maravelakis (2022) investigam o papel do centro de mídia Al-Hayat para propagação da propaganda do Estado Islâmico. No estudo, os autores utilizaram uma combinação de Lexicons e LDA para descobrir tópicos e emoções predominantes nas peças de mídia que o jornal utilizava, sobretudo para propaganda fora do Iraque e da Síria. Os autores obtiveram no estudo os padrões de sentimento nos textos, padrões sistemáticos em tópicos e os tópicos de discussões por documento e agregado.

Quando o tema é comunicação voltada à problemática de Mudanças Climáticas, existem pesquisas como Crawley, Coffé e Chapman (2020) Jiang et al (2017) e Pathak e Volkova (2017) que utilizam de tecnologias similares para obter insights sobre opinião pública, difusão da cobertura midiática e a reação pública por mídia social. Assim, a utilização desse tipo de tecnologia também se mostra relevante para esse tipo de pesquisa.

Em “Public opinion on climate change: Belief and concern, issue salience and support for government action” Crawley, Coffé e Chapman (2020) procuram uma forma de explicar a relação entre opinião pública e mudanças climáticas. Nesse contexto, os

autores coletam dados de pesquisa pública sobre o tema e procuram diferenciar o problema em 3 categorias, quais sejam conhecimento e preocupação, saliência de problemática e suporte para ações do governo. Nessa pesquisa, os autores utilizam uma metodologia chamada Latent Class Analysis (LCA), uma forma de modelar em tópicos usando inferência clássica. Os autores concluem que a grande maioria tem consciência do problema, mas eram cautelosos sobre como o governo pode agir ou não possuíam saliência de tópicos sobre o tema.

Em “Comparing Attitudes to Climate Change in the Media using sentiment analysis based on Latent Dirichlet Allocation”, Jiang et al (2017) estudam como diferentes noticiais abordam o tema da mudança climática, tomando consciência que diferentes agências descrevem a problemática sobre diferentes pontos de vista. Para atingir esse objetivo, os autores coletaram notícias que possuem a palavra-chave “mudanças climáticas” do The Guardian, The Times, The Telegraph e The Independent no período de 2007 a 2016. Para mapear os tópicos e as palavras chaves em tópicos, os autores utilizam uma combinação de LDA com um “Sentiment Lexicon” para obter, analogamente ao estudo anterior, sentimentos predominantes no texto. Como resultado do estudo, os autores concluíram que existem atitudes diferenças significativas nas diferentes fontes de notícias. Por fim, os autores apontam a extensão do estudo para identificar quão sistematicamente essas atitudes variam no tempo.

Em “Understanding Social Media’s Take on Climate Change through Large-Scale Analysis of Targeted Opinions and Emotions”, Pathak, Henry e Volkova (2017) pesquisam como a opinião pública e, consecutivamente, o comportamento social se apresentava em relação ao evento da conferência de mudanças climáticas da ONU em 2015. Nesse estudo, os autores coletaram uma quantidade significativa de contas de Twitter em diferentes lugares do mundo e as organizam em categorias pré-definidas. Por conseguinte, os autores analisam o sentimento predominante por tópico de cada tópico. A pesquisa traz uma integração rica entre modelagem de tópico, mineração de redes sociais e análise de documentos.

Uma terceira perspectiva de análise que utiliza a mesma classe de métodos é como a comunicação institucional do Estado ou de órgãos de organizações internacionais afetam o mercado. Nessa seara, três estudos notáveis são Panagiotis e Andrianos (2017), Edison e Carcel (2021) e Alfredi e Gabrielyan (2021). Os dois primeiros trabalhos investigam o papel da comunicação do Federal Open Market Committee (FOCM) americano e o terceiro o do Banco Central Europeu. Esses trabalhos lançam luz sobre a importância da comunicação para formulação de política pública para o setor.

Em “Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts”, Edison e Carcel (2021) pesquisam sobre a evolução de diferentes

tópicos discutidos por autoridades do Federal Open Market Committee (FOCM) em perspectiva ao acontecimento da Crise Global Financeira de 2008. Utilizando de uma modelagem LDA, os autores analisam mais de 45 mil documentos apresentados pelo comitê. Como resultado, os autores observam que as discussões sobre modelagem econômica se tornam mais intensas no período da crise financeira internacional de 2008 e os sistemas bancários após a crise. Esse estudo é uma importante contribuição para avaliar como a comunicação e as agendas de interesse se manifestam através dos discursos.

Em “Latent semantic analysis of the FOMC statements”, Panagiotis e Andrianos (2017) pesquisam sobre o conteúdo das declarações do FOCM e seu respectivo impacto no mercado americano. Os autores aplicam uma técnica chamada de Latent Semantic Analysis (LSA), um precursor dos modelos LDA, para identificar padrões de associações entre os textos em dados coletados pelo FOCM entre 2003 e 2014. Em seguida, os autores utilizam a regressão linear para identificar os temas mais impactantes na flutuação dos mercados. Como resultado, os autores identificam que os temas são significativos para explicar a variação da produção do tesouro pelos tópicos comunicados pelo FOCM é estatisticamente significativa para 3 meses, 6 meses, 2 anos, 5 anos e 10 anos. Esse estudo é particularmente importante por sugerir que a transparência do pensamento da instituição é importante para gerar ambientes de menor volatilidade.

Em “The Communication Reaction Function of the European Central Bank. An Analysis Using Topic Modelling”, Alfredi e Gabrielyan (2021) investigam a reação do mercado à comunicação do Banco Central Europeu. Para essa investigação científica, os autores utilizam um modelo LDA para construções dos tópicos teóricos e testam sua contribuição para o crescimento do PIB total da EU com defasagem através de uma Função de Reação à Comunicação. Após análise, o estudo apontou uma série de limitações estatísticas relacionadas em relação a pré-requisitos para uma análise linear clássica, mas ainda assim sustenta uma correlação significativa.

Todos os estudos supracitados permeiam o escopo agregada de conteúdo midiático, seja ele de mídias sociais, alternativa ou tradicional. O que há em comum nos estudos interdisciplinares é a abordagem de modelagem estrutural de tópicos, algumas vezes conjuntamente apresentada com análise de sentimentos ou análise de emoções. Essa abordagem é pertinente ao tema por procurar modelar, isto é, explicar estatisticamente através de categorias como o agregado de notícias é difundido. Diferente das pesquisas mais clássicas previamente mencionadas, este trabalho se fundamenta em outros estudos que são mais pertinentes ao campo de Machine Learning e Processamento de Linguagem Natural.

3 Metodologia

Esta pesquisa é baseada em métodos mistos de análise. Do lado qualitativo, o Estudo de Caso é o tipo selecionado para analisar. Para isso, analisa-se o contexto histórico da formação do Fórum Econômico Mundial e apresenta-se uma interpretação teórica sobre a organização baseado em um levantamento bibliográfico narrativo. Do lado quantitativo, coletam-se dados de diversas fontes de notícias e analisa-se através de técnicas de Machine Learning e Deep Learning a difusão de notícias em termos de tópicos e em termos de análise de emoções. Assim, o método qualitativo estabelece um background teórico-analítico sobre o qual a análise objetiva dos dados é feita.

3.1 Revisão de literatura

A construção da fundamentação teórica do Fórum em perspectiva histórica e analítica é realizada através de revisão bibliográfica do tipo narrativa. "O objetivo principal de uma revisão de literatura tradicional ou narrativa é analisar e resumir um corpo de literatura [...] apresentando uma base abrangente da literatura dentro do tópico interessado"(O'GORMAN e MACLINTOSH, 2015, p.37-38, tradução nossa).

A estratégia dessa revisão está relacionada à baixa quantidade de literatura específica sobre a problemática do Fórum Econômico Mundial, no geral, e do papel da mídia no Fórum, em particular. Tomando como base a revisão narrativa, as fontes principais para descrição histórica do Fórum e conceituação são os trabalhos de Pigman (2007), Khanna (2010) e Friesen (2020); documentos oficiais; notícias e discursos escolhidos, como complementares às fontes principais.

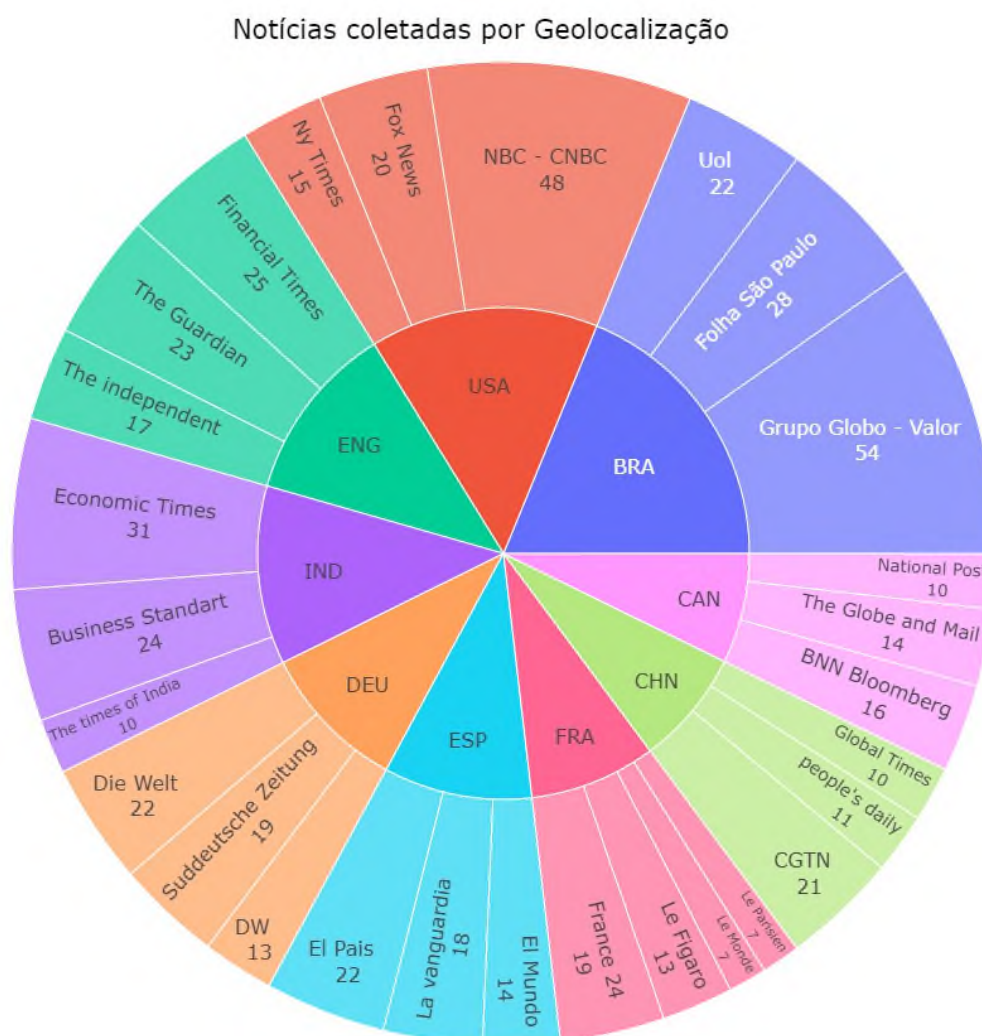
Adicionalmente, apresenta-se também uma revisão bibliográfica semi-sistemática de como a mídia pode ser estudada em diferentes tópicos. Nessa breve revisão, apresenta-se um trabalho sistemático clássico, um quantitativo baseado em estatística clássica e oito outros trabalhos baseados em Machine Learning.

3.2 Obtenção de dados e pre-processamento

O tipo principal de dados obtidos nessa pesquisa é textual, oriundos de notícias públicas em grandes meios de comunicação. A seleção de notícias se deu pelas palavras-chave "Davos" e "Fórum Econômico Mundial" através do navegador do Google, estabelecendo o espaço temporal entre 20 e 30 de maio de 2022, quando

o evento ocorreu. Essas notícias foram coletadas por técnicas de WebScraping buscando preservar todo o conteúdo, com exclusão de elementos não textuais. Os dados coletados são de agências de notícias do Brasil (BRA), Estados Unidos (USA), Canadá (CAN), Inglaterra (ENG), Alemanha (DEU), Espanha (ESP), França (FRA), Índia (IND) e China (CHN). A distribuição de suas frequências pode ser representada na figura 1.

Figura 1 – Gráfico de notícias coletadas por geolocalização



Fonte: Elaboração própria

Partindo do Brasil em sentido anti-horário, obtém-se, em sequência, os países com maior e menor número de notícias coletadas. O critério de seleção das fontes se deu pela disponibilidade de notícias. Nesse contexto, a França possui 4 fontes, devido a baixa disponibilidade.

Obtidos os dados, o processo de tradução para inglês é iniciado pelo API

padrão do Google Translate. Esse procedimento visa construir uma base analítica comum, de tal forma que a mesma palavra em idiomas diferentes não conte como uma palavra nova, mas como uma ocorrência da mesma palavra. Após o estabelecimento de um padrão comum para os dados, realiza-se a metodologia padrão sugerida pelo pacote Gensim da linguagem de programação Python, que consiste nos seguintes procedimentos de tratamento de dados visando a diminuição de dimensionalidade: (1) remoção de símbolos, números e caracteres especiais; (2) remoção de palavras de parada; (3) organização em Bigramas; (4) Lemmatização;

Os dois primeiros procedimentos têm por objetivo remover elementos que atrapalham o treinamento de um modelo. Esses elementos ou por não serem palavras ou por serem palavras comuns não trazem informações relevantes. O terceiro procedimento objetiva identificar conjuntos de duas palavras com alta frequência, como "Joe Biden" e "President Joe" para notícias que tratam do presidente americano. Essa técnica é utilizada visando descobrir padrões que representem combinações de palavras-chave. O quarto procedimento objetiva agrupar e reduzir as formas flexionadas de uma palavra. Ao fazer isso, diminui-se o tamanho do vocabulário e aumenta-se a frequência de palavras que sintetizam toda a complexidade flexional. Após todos os procedimentos, os dados estarão prontos para servirem de input para um modelo de Machine Learning e Deep Learning.

3.3 Modelagem de tópicos

"A modelagem de tópicos é uma técnica que vem com grupo de algoritmos que revelam, descobrem e anotam estrutura temática na coleção de documentos" (KHERNA e BANSAL, p.1, 2019, tradução nossa). Tomando como base esse conceito, podemos considerar algoritmo de forma genérica, isto é, probabilístico, quando baseados em distribuição de probabilidades, e não probabilísticos, quando baseados em modelos algébricos (KHERNA e BANSAL, 2019). Para o caso de análise particular, comparam-se três modelos que pertencem à mesma família de modelos probabilísticos do tipo Latent Dirichlet Allocation.

3.3.1 Os modelos selecionados: LDA, ProdLDA e NeuraLDA

A escolha do LDA está relacionada ao indicativo dos estudos de Chang et al (2009) e Anaya (2011) sobre a boa capacidade desses modelos de se produzir tópicos que se aproximam da compreensão humana. Porém, o escopo dessa pesquisa não ficou limitado apenas ao modelo clássico. Comparativamente a ele, procurou investigar o desempenho de subtipo de modelos LDA baseado em redes neurais. Esses modelos são chamados de ProdLDA e NeuraLDA e foram criados

por Srivastava e Sutton (2017). A justificativa da escolha desses dois modelos está relacionada ao indicativo de performance superior em relação ao LDA clássico (SRIVASTAVA e SUTTON, 2017) e a disponibilização pública pela comunidade.

LDA foi pioneiramente apresentado no campo de Machine Learning por David Blei, Andrew Ng e Michael Jordan no artigo “Latent Dirichlet Allocation” (2003). LDA pode ser definido como “[...] um modelo probabilístico generativo de um corpus [cuja] ideia básica é que os documentos sejam representados como misturas aleatórias sobre tópicos latentes, onde cada tópico é caracterizado por uma distribuição sobre palavras” (BLEI, NG e JORDAN, 2003, p.996).

Em outras palavras, o modelo é alimentado por um conjunto de documentos (d) em que cada documento é representado por um conjunto de palavras (i) que podem ser classificadas em um conjunto de tópicos desconhecidos à priori (k). Se não se conhece o número de tópicos (k), a distribuição de palavras por tópicos (θ_k) e a distribuição de documentos por tópicos (π_d), uma forma de partir do zero é inferi-las através da inferência bayesiana. Baseado nisso, podemos compreender modelo generativo como aquele que objetiva criar documentos novos ($\hat{d}$), que se assemelhem aos documentos já existentes (d) por parâmetros que compõem distribuições de documentos por tópicos e palavras por tópicos (BLEI, NG e JORDAN, 2003).

Segundo Blei, Ng e Jordan (2003), os princípios principais do LDA são “Saco de Palavras” e Permutabilidade. O primeiro princípio é prega que “[...] a ordem das palavras em um documento pode ser desprezada” e o segundo que “[...] ordenação específica dos documentos em um corpus pode ser desprezada” (BLEI, NG e JORDAN, p.994-995, 2003). A construção do modelo também parte das seguintes concepções:

D 1 *Palavras (w) são unidades básicas de dados discretos. Elas ocupam uma única posição em um vocabulário indexado, tal que $w_v = (1, 0, 0, \dots, V)$*

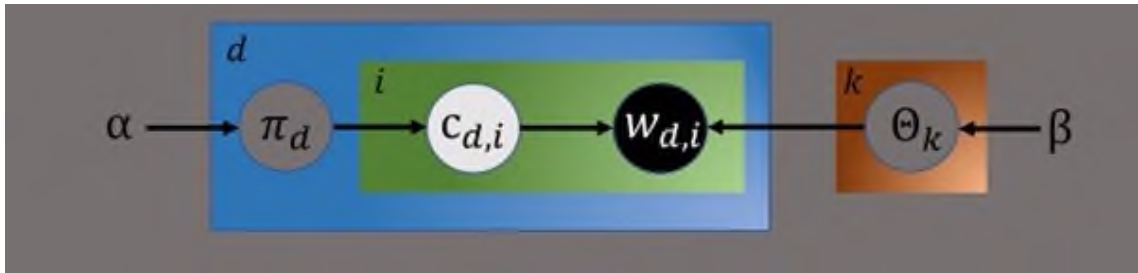
D 2 *Documentos (d) são sequências de N palavras. Assim, $d = (w_1, w_2, w_3, \dots, w_n)$*

D 3 *Corpus (C) é uma coleção de documentos. Assim, $C = (d_1, d_2, d_3, \dots, d_n)$*

Esse modelo pode ser representado pela figura 2. Na figura, visualiza-se uma representação da estatística bayesiana de plate notation, utilizada para indicar repetição em modelos gráficos. Nesse tipo de modelo, a única variável realmente observada são as palavras $w_{d,i}$, que descreve uma palavra no conjunto de palavras de um documento i e de um conjunto de documentos d , ambos representados nos retângulos. Além disso, cada palavra possui um peso em uma distribuição de palavras por tópicos θ_k .

Em síntese, para cada um dos três espaços d , i e k existem parâmetros $\pi_d, C_{d,i}$ e θ_k , respectivamente. Todos esses parâmetros são latentes, isto é, dados não observados que descrevem uma distribuição de probabilidade específica (Hening, 2021). As distribuições de probabilidade obtidas na normalização dos parâmetros seguem uma distribuição de probabilidades do tipo Dirichet para π_d e θ_k e multinomial para $C_{d,i}$. Finalmente, o hiperparâmetros α e β são valores que moldam o formato de cada uma das distribuições Dirichlet.

Figura 2 – Figura do plate notation do LDA



Fonte: Elaboração própria baseada em BLEI, NG e JORDAN (2003)

Antes de adentrar propriamente nas características do modelo e traduzir a figura em linguagem matemática, cabe mencionar a importância da regra de Bayes para estatística bayesiana. Segundo a regra de Bayes, é possível descobrir uma probabilidade condicional de um evento A dado um evento B através da seguinte equação:

$$P(A | B) = \frac{P(A | B)P(A)}{P(B)} \quad (3.1)$$

Os termos $P(A | B)$, $P(A | B)$, $P(A)$ e $P(B)$ são chamados, respectivamente, de probabilidade posteriori, probabilidade à priori, verossimilhança e evidência (ou verossimilhança marginal). A intuição é simples, se compreendermos A como um parâmetro e B como dados, podemos utilizar a regra para descobrir um conjunto de parâmetros que possam representar corretamente o maior número possível de dados. Em outras palavras, o objetivo é maximizar a probabilidade à posteriori, isto é: dados os parâmetros, os dados sejam realmente aqueles. Esse mesmo princípio baseia a compreensão do LDA, mas ao invés de um parâmetro, três são postos para prever documentos. Assim, poderíamos reescrever a equação (3.1) da seguinte forma:

$$P(C, \pi, \theta | w) = \frac{P(w | c, \pi, \theta)P(c, \pi, \theta)}{\int \int \int P(c, \pi, \theta, w)dc d\pi d\theta} \quad (3.2)$$

Segundo Hening (2021), as variáveis são consideradas independentes uma das outras. Logo, os termos superiores da fração podem ser reescritos como na

equação (3.3) e fatorizados como nas equações (3.4) a (3.7)

$$P(w | c, \pi, \theta)P(c, \pi, \theta) = P(\theta | \beta)P(\pi | \alpha)P(c | \pi)P(w | c, \theta) \quad (3.3)$$

sendo ,

$$P(\theta | \beta) = \prod_{K=1}^K D(\theta, \beta) \quad (3.4)$$

$$P(\pi | \alpha) = \prod_{D=1}^D D(\pi, \alpha) \quad (3.5)$$

$$P(c | \pi) = \prod_{D=1}^D \prod_{I=1}^I \left(\prod_{K=1}^K \pi_{d,k}^{c_{d,i,k}} \right) \quad (3.6)$$

$$P(w | c, \theta) = \prod_{D=1}^D \prod_{I=1}^I \left(\prod_{K=1}^K \theta_{d,k}^{c_{d,i,k}} \right) \quad (3.7)$$

Observa-se no lado esquerdo da equação (3.2) que se quer descobrir um conjunto de parâmetros dado novas unidades de dados w . À medida que novas palavras são inseridas, elas ganham novos parâmetros explicativos. O agregado dos parâmetros, que representa cada uma das equações (3.4) a (3.7), são obtidos através de distribuições de probabilidades. As equações (3.4) e (3.5) representam a distribuição de Dirichlet obtidas pelos hiperparâmetros α e β , respectivamente. Essas distribuições do tipo Dirichlet é conhecida por ser uma "distribuição de distribuições". Assim, dada uma distribuição de Dirichlet, será escolhida a distribuição multinomial que consiga aproximar ao máximo a condição $\hat{d} = d$.

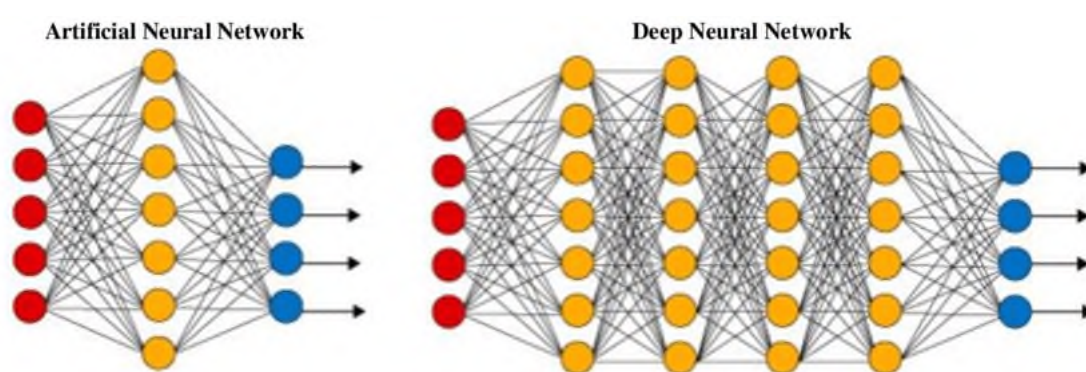
NeuraLDA é basicamente a aplicação do algoritmo LDA transferindo o aprendizado do processo para um modelo de Deep Learning. Esse tipo de modelo é baseado em Redes Neurais Artificiais (RNA), criado por Rumelhart, Hinton e Williams (1986), e é muito útil para modelos generativos. Basicamente, os autores consideram que o processo de aprendizagem de máquina pode se assemelhar à estrutura do aprendizado humano, de tal forma que existam unidades que recebem impulsos e repassam outros impulsos para outros neurônios, desencadeando uma resposta. Por analogia, o olho humano pode observar um pássaro e um encadeamento de impulsos neuronais é lançada para interpretar aquilo de alguma forma. A rede neural ao "ver" um pássaro pode produzir uma foto de outro pássaro, ou descrever sua nomenclatura após aprender em um dataset o que é um pássaro. Basicamente, o modelo procura o melhor caminho para que o output se assemelhe ao máximo ao que é o dado real.

No caso específico, NeuraLDA e ProdLDA são redes neurais artificiais do tipo de Aprendizagem Profunda. Isso significa dizer que toda a parte inferencial é feita "como um cérebro humano", em que neurônios divididos em camadas ocultas produzem impulsos (funções de ativação) que cada um dos neurônios executa

(SRIVASTAVA e SUTTON, 2017). Ao receber os dados, a RNA calcula os parâmetros em funções de ativação e produz um documento $\hat{d}$ que erre em maior ou menor grau o conjunto de dados. À medida que o modelo segue errando na predição, ele ajusta os parâmetros, descobrindo uma forma de classificar melhor as informações. Esse fenômeno de aprender com os erros foi definido por Rumelhart, Hinton e Williams (1986) como retropropagação.

Para elucidar esses passos, a figura 3 a seguir representa a Estrutura de uma rede neural artificial de aprendizagem profunda.

Figura 3 – Figura de Redes Neurais Artificiais



Fonte: Mostafa et al 2020

Na figura, existem dois tipos de redes neurais artificiais. A primeira é uma RNA genérica e a segunda é uma de RNA de aprendizagem profunda. Cada coluna de círculos é chamada de camada. A primeira camada é chamada de Entrada, em que os pontos vermelhos são os inputs, isto é, o corpus do modelo. O conteúdo desse corpus será decomposto em camadas ocultas, que representam cada uma das colunas com pontos amarelos, sendo cada ponto amarelo um neurônio ¹. A última camada produz um conjunto de documentos (outputs). No fim, são computados os gradientes da função de perda de cada um dos elementos treinados. Com os gradientes calculados, utiliza-se uma técnica de Machine Learning chamada de Gradiente Descendente para achar a melhor combinação de parâmetros para cada elemento.

Nesse contexto, ProdLDA a principal diferença entre NeuraLDA e da LDA é apenas o método de distribuição de palavras por tópicos ($C_{d,i,k}$). Ao invés dessa distribuição ser multinomial, ela segue um algoritmo de Machine Learning chamado de produto das expectativas para se descobrir a distribuição de probabilidades, mantendo

¹ "Um neurônio artificial, é uma função matemática que leva uma ou mais entradas para obter uma função linear que é passada para uma função não linear, também conhecida como função de ativação para obter a saída do neurônio"(PRABHU, 2020, p.1, tradução nossa)

as distribuições de Dirichlet para $\pi e \theta$. Matematicamente, a distribuição C pode ser computada através da proporcionalidade ($\propto$) descrita na equação abaixo

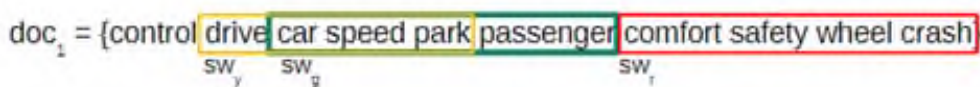
$$P(w | C, \beta) \propto \prod_{K=1}^K p(w_n | z_n = K, \beta)^{C_k} \quad (3.8)$$

3.3.2 Problema da otimização, métrica de qualidade e nomenclatura de tópicos

Dada a consciência do pré-processamento e do funcionamento dos modelos, inicia-se o problema de avaliação e otimização, que conjuntamente constituem o esforço científico de diferenciar bons e más modelos mudando as condições de treinamento. Partindo da problemática da avaliação, em "Topic Coherence Metrics: How sensitive are they ?", Campagnolo et al (2022) abordam a complexidade de se rotular uma métrica de qualidade para modelos não supervisionados, por ainda não existir uma métrica totalmente precisa para avaliação. Paralelo a isso, em "Words alone: Dismantling topic models in the humanities", Schmidt (2012) defende que modelos possuem duas propriedades fundamentais: coerência e estabilidade.

Coerência (RÖDER, BOTH e HINNEBURG, 2015) é uma métrica de qualidade baseada em outras métricas antecessores (Chang et al, 2009; Lau et al, 2014) que visam explicar como se dá a co-ocorrência de palavras (CAMPAGNOLO et al, 2022). Ambos trabalhos de Röder, Both e Hinnerburg (2015) e CAMPAGNOLO et al (2022) para Coerência, e Chang et al (2009) e Lau et al (2014) para outras métricas próximas à coerência, apontam uma forte correlação entre a interpretabilidade de máquina e a interpretabilidade humana. Em especial, ambos Röder, Both e Hinnerburg (2015) e Campagnolo et al (2022) apontam que a coerência tipo C_v e C_p são as que melhor performam na interpretabilidade humana, com resultados bastante parecidos. Devido à falta de acessibilidade à métrica C_p na comunidade do python, utiliza-se apenas a C_v . A intuição de ambas as métricas é descobrir a frequência da co-ocorrência de palavras em uma janela de palavras que procedem uma determinada palavra centralizada. A figura 4 representa essa ideia intuitiva.

Figura 4 – Sliding Window



Fonte: CAMPAGNOLO et al (2022)

Sob uma janela de 110 palavras cria-se múltiplos vetores representando a co-ocorrência de W em comparação com as outras. Assim, pode-se representar

matematicamente o vetor

$$v_{W',W^*} = NPMI(W', W^*)^\gamma \quad (3.9)$$

sendo NPMI a normalização do PMI, uma métrica para associação entre duas palavras. Existe um segundo vetor que reúne todos os vetores da equação acima, definido por

$$\vec{v}_i = \{NPMI(W', W_1^*)^\gamma, \dots, NPMI(W', W_n^*)^\gamma\} \quad (3.10)$$

A coerência é então calculada pela média da similaridade do conseno, através da fórmula

$$C_v = \frac{(\cos(\vec{v}_i, \vec{v}_c) + \dots + \cos(\vec{v}_n, \vec{v}_c))}{N} \quad (3.11)$$

sendo cos a similaridade dos cosenos e N o número total

Além da coerência existem métricas mais antigas como Perplexidade, introduzido por Grun e Hornik (2011), com vasta aplicação científica para modelos de ML (Newman et al, 2009; Du e Jin, 2010; Zhai e Boyd-Graber, 2013; Zhao et al, 2017) e Deep Learning (Ding, Nallapati e Xiang, 2018; Gui et al, 2019; Razaee e Ferraro, 2020). Essa métrica tange ao segundo aspecto fundamental de um modelo de modelagem de tópicos: a estabilidade (ZAHO et al, 2015). "Ele captura o quão surpreso um modelo fica com novos dados que não viu antes e é medido como a verossimilhança logarítmica normalizada de um conjunto de teste mantido." (KAPADIA, 2019). Porém, como sustenta Zhao et al (2015) em "A heuristic approach to determine an appropriate number of topics in topic modeling" essa métrica é instável e pode ser muito sensível perturbações durante as interações. Por conseguinte, Zhao propõe como método adequado de se achar o número adequado de tópicos a observância da taxa de variação de perplexidade, definida pela fórmula

$$RTPC = \left| \frac{\Delta Pe}{\Delta K} \right| \quad (3.12)$$

Sendo perplexidade definida como:

$$Pe = perplexity(C) = \exp\left\{-\frac{\sum_{d=1}^C \log p(w_d)}{\sum_{d=1}^C N_d}\right\} \quad (3.13)$$

em que N é o tamanho do vocabulário, w_d é uma palavra para cada documento e d é um documento do Corpus (C) .

A coerência é a métrica principal no que tange à interpretabilidade humana, e por isso será considerada a principal, mas para diferenciar uma seleção de modelos, adotara-se o método heurístico de Zhao et al (2015). Enquanto o objetivo da primeira métrica é a maximização (RÖDER, BOTH e HINNEBURG, 2015), o da segunda é

escolher e avaliar os menores pontos que satisfaça a condição $RTPC_t < RTPC_{t-1}$ (ZHAO et al, 2015).

Porém, como bem aponta Campagnolo et al (2022), essas métricas são guias, mas não determinísticas. Isso significa que uma análise subjetiva é necessária sobre os resultados filtrados, levando em consideração o comportamento dos tópicos ao se alterar sua quantidade. Em outras palavras, busca-se compreender se uma parte dos tópicos se conserva e se sua alteração favorece uma análise subjetiva do problema. Assim, procura-se olhar se as 5 notícias mais representativas produzidas por cada tópico ajudam a ganhar insights sobre um tópico.

Após definida a métrica de análise, o problema da otimização é lançado. No total, cada tipo de modelo passa por 17400 simulações, totalizando 52200 para os 3 modelos. Dessas 17400 simulações, cerca de 600 provém de cada K, sendo 60 combinações de hiperparâmetros diferentes repetidos 10 vezes a fim de evitar falta de confiança dos resultados. Para cada combinação, produz-se um modelo, recolhe suas métricas de qualidade e se testa um novo modelo com uma nova combinação de hiperparâmetros, seguindo a própria metodologia proposta por Terragni e Fersini (2021). As tabelas 1 e 2 a seguir descrevem os hiperparâmetros para o modelo LDA e os modelos NeuraLDA e ProdLDA.

Tabela 1 – Tabela dos hiperparâmetros para LDA

Métrica	Espaço de busca	Tipo de Intervalo
α	$[10^{-3}, 5]$	Contínuo
β	$[10^{-3}, 5]$	Contínuo
TW	[ONE, IDF, PMI]	Contínuos
K	[2, 30]	Discreto

Fonte: Elaboração própria com base em Terragni e Fersini (2021) e Wilson e Chew (2010)

Tabela 2 – Tabela dos hiperparâmetros para ProdLDA e NeuraLDA

Métrica	Espaço de busca	Tipo de Intervalo
Dropout	[0, 0.9]	Contínuo
Momentum	[0, 0.9]	Contínuo
Número de Neurônios	[100, 1000]	Discreto
Número de Camadas	[1, 5]	Discreto
K	[2, 30]	Discreto

Fonte: Elaboração própria com base em Terragni e Fersini (2021)

α , β e K já foram introduzidos na seção anterior e descrevem os hiperparâmetros para distribuições do tipo Dirichlet. Tw (Term Weight) é uma adaptação feita por Wilson e Chew (2010) para LDA que descreve a metodologia utilizada para distribuir as palavras mais recorrentes em tópicos.

Dropout, Momentum, Número de Neurônios e Número de Layers foram os principais hiperparâmetros de redes neurais de aprendizado profundo. Os dois últimos foram apresentados na subseção anterior. Dropout controla o descarte de neurônios no processo de aprendizagem, ignorando seu peso no modelo visando reduzir o problema de sobreajuste² de um modelo (BROWNLEE, 2018). Momentum controla a retenção do gradiente para dar mais velocidade no treinamento e contribuir com o processo de otimização.

À rigor, o processo de nomenclatura de um tópico é um processo subjetivo (ALETRAS et al, 2014). Apesar de haverem metodologias de nomenclatura automática (MEI, SHEN e ZHAI, 2007; SNYDER et al, 2013), essas metodologias muitas vezes são contra-intuitivas na sintetização do Corpus. Aletras et al (2014) e Lau et al (2010) propõe olhar para as palavras com maior probabilidade marginal a fim de delimitar o espaço possível de nomes, ou a seleção de um único nome que sintetize essas palavras. Baseado nessa visão, e tomando como base a subjetividade da seleção de um nome que sintetize o tópico, este trabalho utiliza de um artifício novo para delimitar a nomenclatura subjetiva.

Ao invés de olhar as palavras com maior probabilidade marginal, observa-se a ocorrência dessas palavras nos 20 documentos mais representativos de cada tópico. Ao fazer isso, procura sintetizar a complexidade do conteúdo de diferentes notícias a partir de exemplos significativos, isto é, exemplos que contemplem a maior parte das palavras significativas para cada tópico. Ao olhar para ocorrência dessas palavras, procura-se observar a frequência que elas ocorrem nos melhores documentos de cada tópico, sintetizando a complexidade do nome que um tópico pode levar.

3.4 Análise de emoções por redes neurais

Paralelamente à modelagem de tópicos, utiliza-se de Deep Learning também para se computar emoções nos textos originais - isto é, sem o pre-processamento a partir de interpretação de máquina. O algoritmo selecionado é baseado no Bidirectional Encoder Representations from Transformers (BERT), disponibilizado publicamente pela Google à comunidade de Machine Learning em 2018. Desde então esse tipo de modelo vem servindo para tradução, visão de máquina, sumarização de textos e detecção de emoções e sentimentos. O modelo selecionado é especificamente "bert-base-uncased-emotion", disponibilizado publicamente por

² Esse conceito se refere ao problema de quando um modelo se ajusta bem aos dados treinados, mas não é muito eficaz em prever bons resultados para novos dados (BROWNLEE, 2018)

Devlin et al (2018). Usando F1-score como métrica do modelo ³, ele apresentou sucesso superior à Liu et al (2019), Lan et al (2019) e Sanh et al (2020).

O uso das Redes Neurais do tipo Transformers, contidas no BERT, são consideradas uma revolução em Deep Learning. Desenvolvida pela equipe da Google, o modelo que inicialmente servia para tradução foi treinado com 45TB de dados da internet, algo revolucionário para 2018 em relação a outros tipos de modelo de Deep Learning como GPT-3 (MARKOVITZ, 2022). Segundo Markotiz (2022), o modelo é fundamentalmente construído em cima de três princípios: encodificação posicional, atenção e auto-atenção.

A encodificação posicional é uma indexação de cada palavra que represente sua posição em uma frase. Nesse tipo de modelo, não se aplica o princípio do saco de palavras referido na subseção 1.2.2, isto é, a ordem das palavras importa. Após esse passo, é calculada uma função de ativação chamada de atenção (VASWANI et al 2017). Ao treinar com Big Data, o modelo vai criando representações internas capazes de prever corretamente alguma tarefa. A inovação verdadeira, desenvolvida seminalmente por Vaswani et al (2017) é a auto-atenção. Após criar um modelo de representação próprio, o modelo consegue ser generalizado para outros tipos de atenção, como identificar gramaticalmente parte de discursos e visão computacional.

No caso específico, essa Rede Neural do tipo BERT aprende a detectar emoções em 4 grandes bases de Big Data: Swag (ZELLERS et al, 2018), para aprender sobre inferência do senso-comum; Squad V1.0 (RAJPURKAR et al, 2016) e Squad V2.0 (RAJPURKAR et al, 2018), para treinar sobre habilidades de responder perguntas com respostas curtas e longas, respectivamente; e GLUE (WANG et al, 2018), para múltiplas tarefas de entendimento de linguagem. Treinado sobre cada ambiente, o modelo se torna capaz de detectar elementos que, em contexto, produzem emoções como alegria, amor, tristeza, raiva, surpresa e medo. Nesse contexto, cada documento é considerado uma distribuição de emoções, obtendo uma predominante.

³ Compreendendo-se precisão como a métrica que procura "em tudo o que foi previsto como positivo [...] a porcentagem que está correta" e recall "dentro de tudo o que realmente é positivo, quantos o modelo conseguiu encontrar"(KORSTANGE, 2021, p.1), F-1 pode ser compreendido como a média harmônica dessas métricas, visto que existe um trade-off entre elas.

4 Exploração Analítica da pesquisa

Após a experimentação computacional, os resultados podem ser discutidos nas subseções: (4.1) a seleção entre tipos de modelos de tópicos diferentes e a escolha do número de tópicos, (4.2) Resultados da modelagem de tópicos; (4.3) Resultados da modelagem de emoções.

4.1 A seleção entre tipos de modelos de tópicos diferentes e a escolha do número de tópicos

Partindo da comparação entre diferentes tipos de modelos, a figura 5 sintetiza a diferença entre eles. Os três gráficos na linha superior descrevem cada uma das 17400 simulações para 30 tópicos. Cada ponto é a métrica de C_v para cada média de 10 repetições do modelo. Observa-se que LDA performou bem melhor do que NeuralDA e ProdLDA, produzindo modelos com coerência superior a 0.7 entre 6 e 10 tópicos e 18 e 20 tópicos.

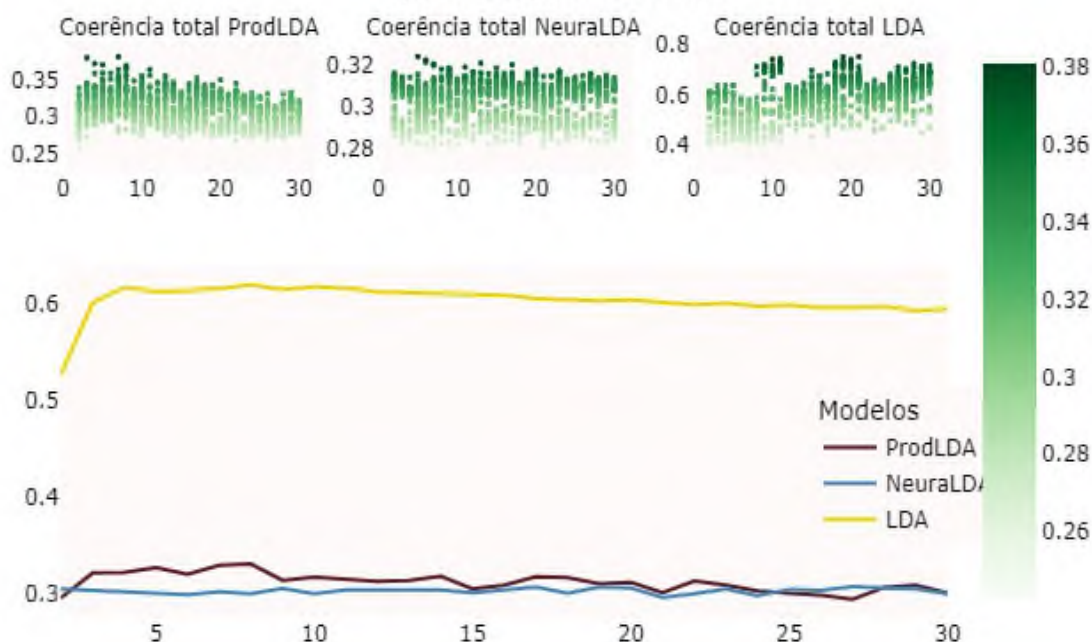
No gráfico 5, traça-se o valor médio para cada tópico dos gráficos superiores, representando cada um das linhas coloridas da legenda, com exceção da branca que é uma terceira média sobre as médias para separar dentre elas aquelas com comportamento acima do ponto médio. Nota-se, como no gráfico anterior, um saliência acima da terceira média, criando-se assim um espaço de investigação para o método heurístico de análise, representado na figura 6.

Na figura, observa-se através dos pontos vermelhos a satisfação da condição $RTPC(i) < RTPC(i - 1)$, sendo os pontos mínimos, representado pela cor verde, os tópicos 9, 5 e 12, representados aqueles com menor RTPC. Essas três sugestões são levadas em conta para a seleção subjetiva do modelo final.

Após duas seções de filtro, uma sobre a coerência e a segunda sobre o RTPC, 3 modelos foram obtidos para análise subjetiva. A partir desse momento, cada um desses modelos será decomposto em tópicos, em que cada tópico representa um gráfico da figura, onde o eixo X representa as palavras mais importantes de cada tópico e o eixo y as probabilidades. As barras coloridas indicam as probabilidades condicionais para cada tópico específico e as barras pretas a probabilidade marginal para aquele tópico. Nesse contexto, os modelos 9 e 12 possuem, na legenda, o tópico originado a partir do modelo de 5 tópicos. Assim, entende-se "K1 from K0*", por exemplo, como o tópico 1 do modelo seguinte provavelmente originado do tópico 0 do modelo 5. Ademais, cada ponto branco é apenas um artefato para indicar que aquela

Figura 5 – Gráfico comparativo entre os modelos ProdLDA, NeuraLDA e LDA

Gráfico de Coerência (Cv) das experimentações de LDA e subvariações para Corpus reduzido



Fonte: elaboração própria com base nos dados obtidos da experimentação

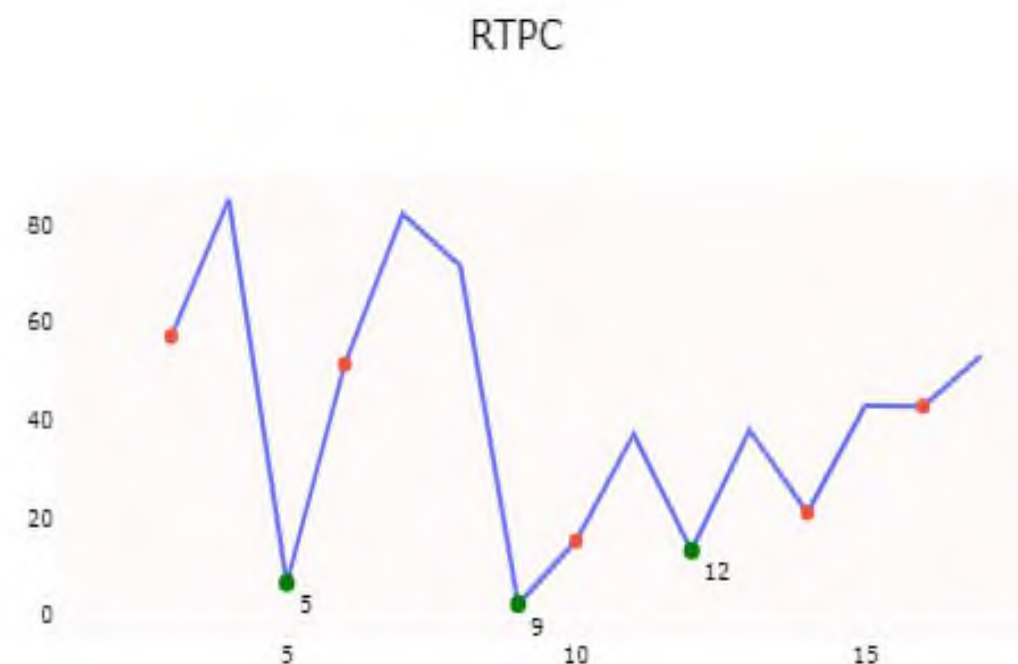
palavra é nova no novo modelo.

Ao observar a primeira distribuição na figura 7, observa-se em cada gráfico de barras cada um dos 5 tópicos do modelo 5. À priori, pode-se abstrair o tópico 0 como aquele ligado à economia e energia, por combinar "inflation", "growth", "price", "market", "recession" e "crisis" com "energy", "oil" e "gas".

O tópico 1 pode ser abstraído como aquele que trata da guerra Rússia e Ucrânia, com as palavras "Russian", "Ukranian", "military", "weapon", "sanction", "attack", "force" e "kyev" e posusi o elemento "grain" como um elemento a ser explorado. O tópico 2 pode ser abstraído como aquele que trata mais especificamente do próprio Fórum, com as palavras "wef", "event", "attend", "meeting", "davo" e "summit" palavras que denotam sentido específico do evento de Davos e do Fórum; "executive" e "party" como palavras mais representativas de política; e "climate" como um ponto a ser explorado.

O tópico 3 pode ser compreendido como aquele que está ligado a mercados, tendo como principais evidências dessa impressão as palavras "investment", "industry", "semicondutor", "criptocurrency", "chip" e "digital"; semanticamente ligadas com outras palavras como "state", "project", "technology" e "sector". O tópico 4 aparenta estar misturado entre saúde e riqueza, com "vaccine", "covid", "pfizer" e

Figura 6 – Gráfico do RTPC para LDA



Fonte: elaboração própria com base nos dados obtidos da experimentação

"health" representando saúde e "tax", "billionaire", "wealth", "rich" e "oxfam" podendo representar riquezas.

O saldo geral do modelo pode ser sintetizado na tabela 3. À priori, os tópicos ainda aparentam estar misturados, gerando baixa especificidade. Por conta disso, compara-se a produção do resultado do modelo 5 com a do modelo 9 na figura 8.

Tabela 3 – Síntese do modelo 5

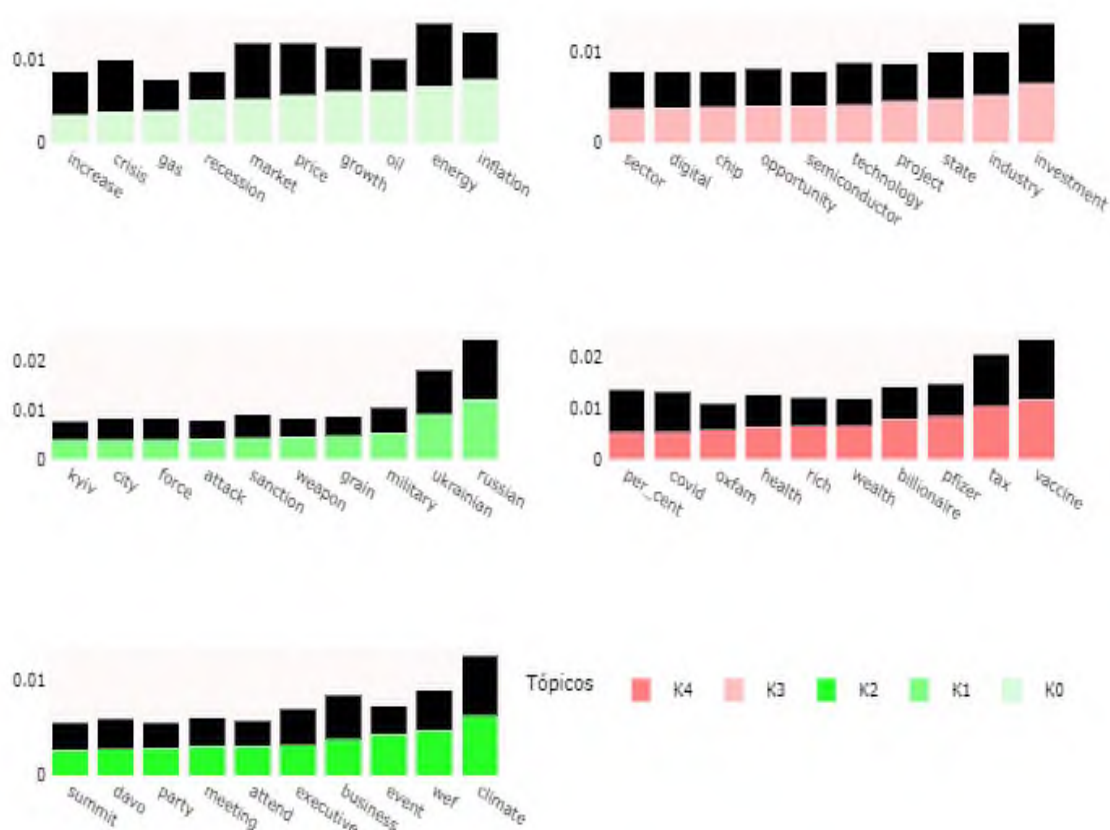
Tópico	Provável conceito
0	Energia e Economia
1	Guerra Rússia e Ucrânia
2	Fórum
3	Mercados
4	Saúde e Riqueza

Fonte: elaboração própria com base nos dados obtidos do modelo 5

Na figura 8, observam-se nove tópicos. Os tópicos 0, 2, 3 e 5 apresentam no mínimo 60% dos tópicos originais no modelo 5. Por conseguinte, o tópico: 0 pode ser entendido como aquele ligado a energia e economia, por combinar 8 palavras do mesmo tópico no modelo 5; 2 pode ser entendido como aquele que trata da "Guerra Rússia e Ucrânia"; 3 pode ser entendido como aquele que trata de "mercados"; e 5

Figura 7 – Distribuição de probabilidade condicional de palavras por tópico (modelo 5 tópicos)

Distribuição de probabilidade condicional e marginal dos termos por tópicos do modelo 5



Fonte: elaboração própria com base nos dados obtidos da experimentação

como aquele que trata do "Fórum". O tópico 1 e 6 aparenta ser uma decomposição do tópico 4 do modelo 5, sendo 1 mais relacionado a saúde e 6 mais relacionado a riqueza.

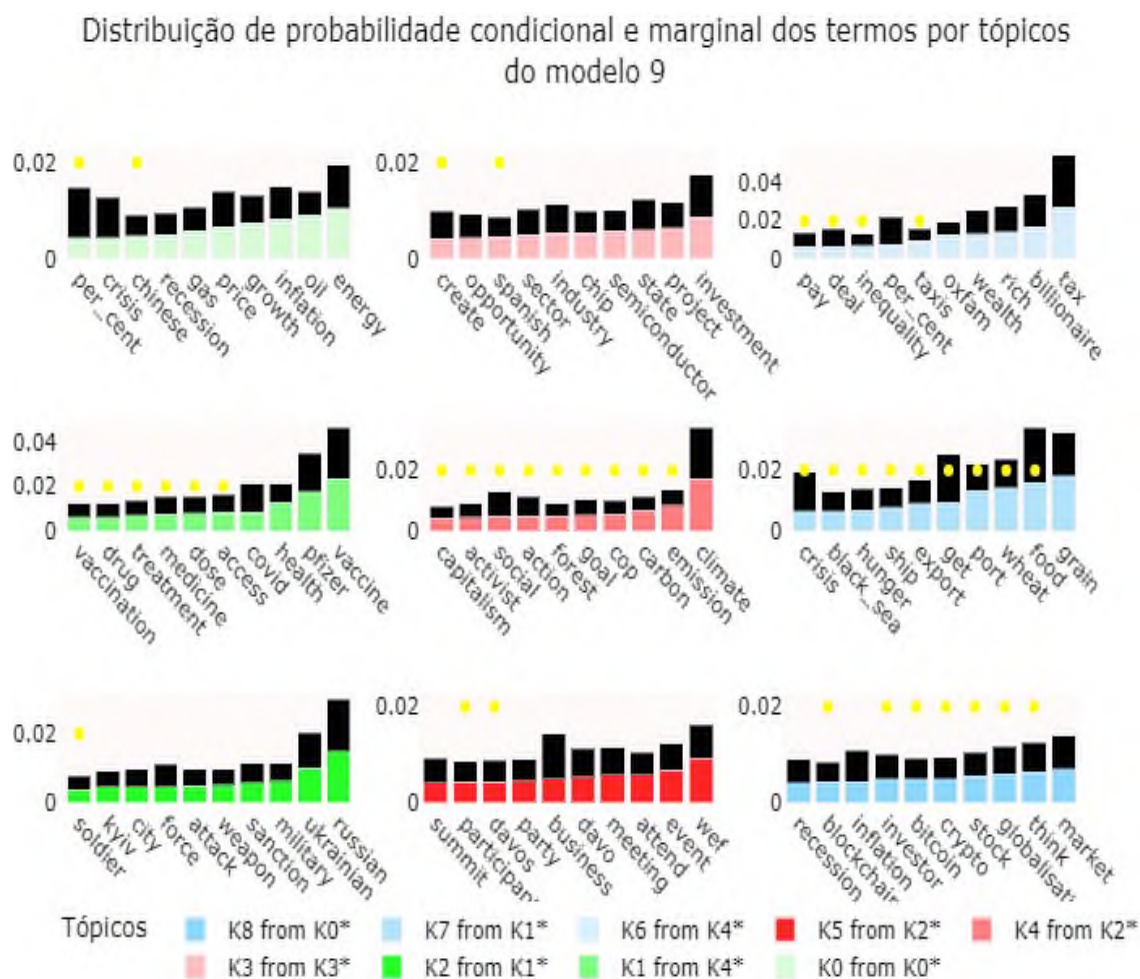
Diferentemente dos tópicos 0,1,2,3,5 e 6, os tópicos 4, 7 e 8 provavelmente remetem à novos tópicos. O tópico 4, aparenta ser um novo tópico a partir do tópico "Fórum" do modelo 5. Seu conteúdo aparenta estar ligado à Clima, pois conta com palavras como "Climate", "Emission", "Carbon", "Cop" e "Forest" e possíveis palavras relacionadas como "goal", "action", "social" e "activist".

O tópico 7 aparenta abrir também uma nova compreensão a partir do tópico de Guerra da Ucrânia, porém possivelmente relacionada ao aspecto da fome, pois apresenta as palavras "food", "grain" e "hunger" indicando fome; "black sea" possivelmente indicando o aspectos específicos provenientes da região e "ship" e "export" relacionados a algum fato específico; e "crisis", possivelmente relacionada

com o contexto ou a um fato específico.

Por fim, o tópico 8 também aparenta ter sido herdado do tópico de energia e economia do modelo 5, mas com a intrusão de 70% de novas palavras relacionadas ao mercado financeiro e moedas digitais como "blockchain", "crypto", "bitcoin" e "stock"; e palavras mais gerais como "investor", "globalization" e "think".

Figura 8 – Distribuição de probabilidade condicional de palavras por tópico (modelo 9 tópicos)



Fonte: elaboração própria com base nos dados obtidos da experimentação

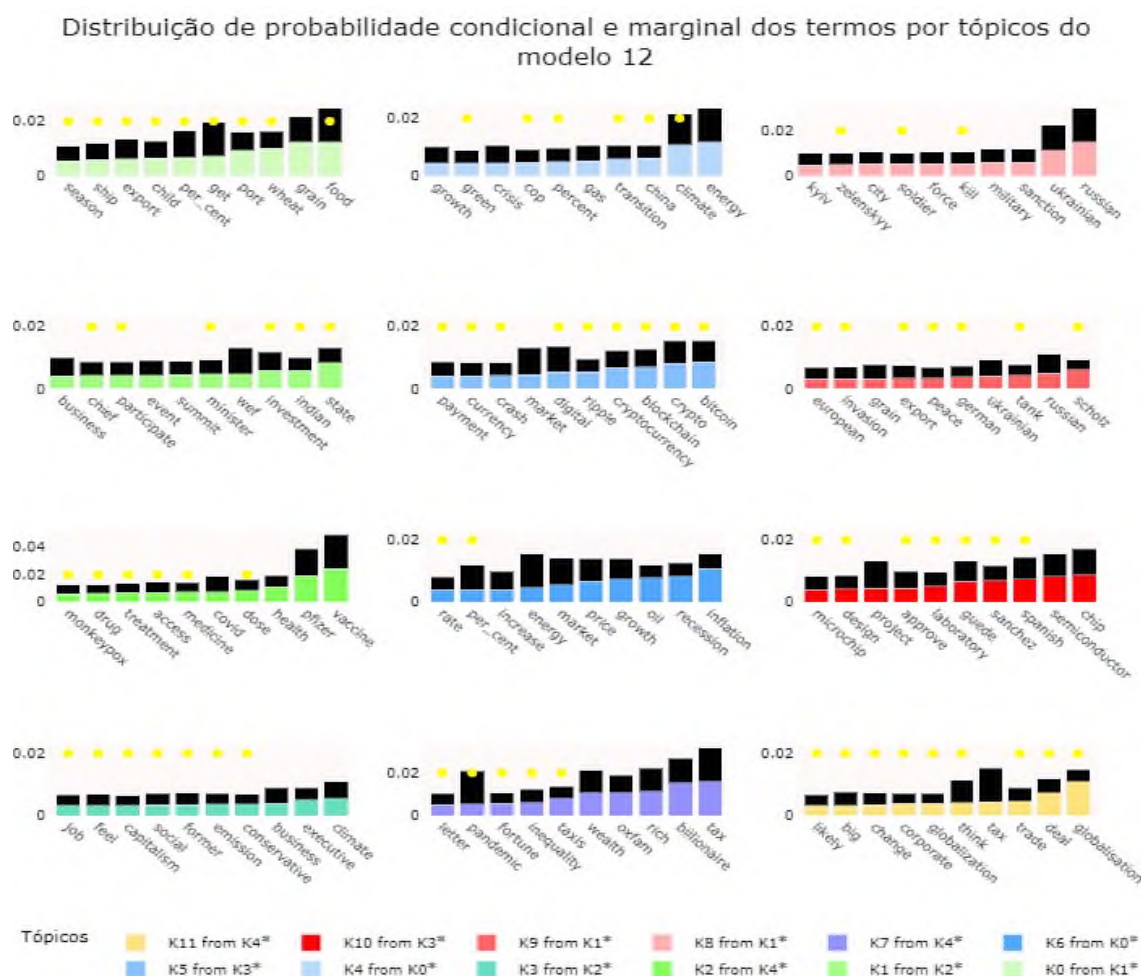
Ao aumentar o número de tópicos de 5 para 9, observa-se que os tópicos 4, 7 e 8 se diferenciam significativamente do modelo com 5, enquanto 1 e 6 aparentam ser uma decomposição com alta similaridade do tópico de saúde e riqueza. Assim, o aumento do número de tópicos pode ter facilitado o entendimento de cada tópico.

Finalmente, a figura 9 trata dos resultados obtidos a partir do modelo com 12 tópicos. O tópico 6 introduz apenas "per cent" e "rate", que, no contexto, não aparenta trazer algum tipo de informação relevante. Similarmente, o tópico 4 aparenta vir de "Energia e Economia", mas introduzindo as palavras "climate", "transition", "cop" e

"green" como evidências de especificidade para o clima.

De outro lado, o tópico 7 aparenta vir do tópico de "Saúde e Riqueza", introduzindo "taxi", "inequality", "fortune" sem prejuízo à compreensão. Similarmente, o tópico 2 também aparenta vir de "Saúde e Riqueza", porém mais voltado a saúde, com introdução de palavras como "dose", "medicine", "access", "treatment", "drug" e "monkeypox".

Figura 9 – Distribuição de probabilidade condicional de palavras por tópico (modelo 12 tópicos)



Fonte: elaboração própria com base nos dados obtidos da experimentação

Os tópicos 1 e 3 parecem ser uma decomposição do tópico "Fórum". O tópico 1 introduz "state", "investment", "indian", "minister" e "chief" que podem indicar um olhar mais voltado ao papel do Estado no evento. Por sua vez, o tópico 3 aparenta trazer também o lado do clima, mas mais voltado aos acontecimentos específicos do Fórum.

Os tópicos 5 e 10, por sua vez, são originados do tópico "Mercados". O tópico 5 aparenta tratar do mercado digital, selecionando palavras como "bitcoin", "cripto", "blockchain", "cryptocurrency", "ripple", "digital" e "currency". O tópico 10 aparenta

tratar mais especificamente do mercado de microchip e semicondutores, com as palavras "chip", "semicondutor", "laboratory" e "microchip", relacionados diretamente ao tópico e "spanish", "sanchez" e "guede" indiretamente.

Os tópicos 8 e 9 aparentam provir do tópico da Guerra Rússia e Ucrânia. O tópico 8 se assemelha muito ao tópico 1 do modelo 5, mas introduz "Kill", "soldier" e "zelensky" sem prejuízo de sentido. Já o tópico 9 aproveita "grain", "russian" e "ukranian", mas introduz outras palavras que podem indicar o contexto específico que as anteriores foram empregadas.

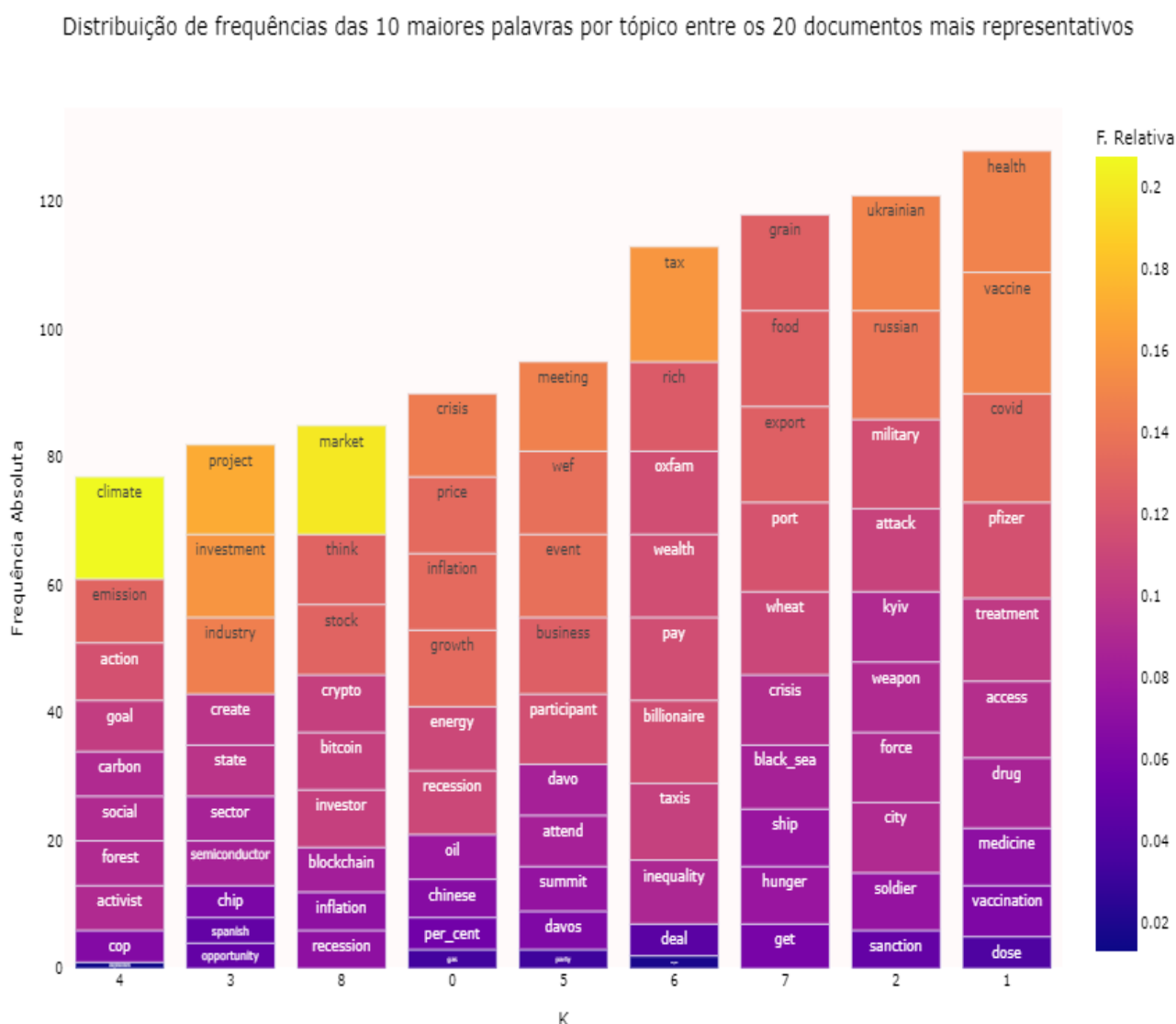
Finalmente, os tópicos 0 e 11 aparentam ser tópicos totalmente novos. Enquanto o 0 mostra o mesmo comportamento do tópico 7 do modelo 9 - a formação de um tópico relacionado a fome - o tópico 11 tem uma relação aparente com negócios internacionais, devido à palavras como "deal", "trade", "globalization" e "corporate", mas apresentam palavras como "likely", "big" e "think" que não aparentam ter especificidade alguma.

Baseado no exposto, o autor identifica subjetivamente que o modelo 5 aparenta ser bastante restritivo, juntando temas como "Saúde e Riqueza" e "Energia e Economia" que podem ser decompostos em outros tópicos, como aponta a transição do modelo de 5 para 9 tópicos. No entanto, ao aumentar de 5 para 12, muitos tópicos se tornam repetitivos ou com probabilidades marginal e condicional baixas, indicando tópicos muito específicos que poderiam ser explicados em categorias maiores. São exemplos desses tópicos 1 e 3, 5 e 10 e 8 e 9. Assim, o aumento dos tópicos não ajuda, à priori, a ganhar informações novas sobre o corpus.

Após as impressões iniciais, cabe analisar mais aprofundadamente a ligação entre as palavras mais representativas e os documentos mais representativos de cada tópico. Para isso, a figura 10 traz uma síntese de como se comportam essas 10 palavras nos 20 documentos mais representativos de cada tópico. Esse comportamento é medido através das frequências por documento, isto é, em quantos documentos a palavra aparece pelo menos uma vez. Assim, o tamanho da barra representa a soma das frequências unitárias de cada uma das palavras por cada um dos documentos, a cor mede a frequência relativa de cada uma das palavras pelo total das frequências, o eixo X representa o tópico observado e o eixo y é a escala de frequências absoluta.

Pela figura 10, observa-se que os tópicos 1, 2, 7 e 6 são aqueles com maior frequência, seguidos de 5, 0, 8, 3 e 4. Também, observa-se uma variação da amplitude de cores por tópicos, indicando que a distribuição de frequências relativas não é igual, podendo haver palavras que sintetizem melhor o conjunto dos melhores documentos por tópicos.

Figura 10 – Distribuição de frequência das 10 maiores palavras por tópico entre os 20 documentos mais representativos



Fonte: elaboração própria com base nos dados obtidos da experimentação

Nesse sentido, os tópicos 4, 8 e 3 são aqueles em que a palavra mais frequente representa por volta de 20% do total, enquanto os demais não passam de 17%. No tópico 4, a soma das 4 primeiras palavras compreende mais de 50% das frequências, sendo "climate" o principal representante, seguido de "emission", "action" e "goal". O conjunto dessas palavras representará a semântica que o tópico apresenta, podendo ser nomeado subjetivamente como "Clima". Por sua vez, o tópico 8 concentra nas palavras "market", "stock", "think" e "bitcoin" mais de 50% das frequências. Nesse sentido, o sentido dado ao tópico 3 é subjetivamente "Mercado Financeiro". Finalmente, o tópico 3 projeta também nas 4 primeiras palavras a concentração semântica de 50%, tendo as palavras "project", "investment", "industry" e "create" como as mais representativas. Por conta dessas características, o tópico

chamara-se "projetos e investimentos".

Para nenhum dos tópicos 0, 5, 6, 7 e 2 as 4 primeiras palavras representam 50% das frequências relativas, representando uma maior distribuição relativa entre as principais palavras. Nesse contexto, o tópico 0 concentra em "crisis", "price", "inflation" e "growth", dando uma semântica subjetivamente identificada como "Economia" de forma mais geral, e "energia" de forma mais marginal por contar com as palavras "energy", "oil" e "gas" com menor frequência. O tópico 5 apresenta as palavras "meeting", "WEF", "event", "business", "partipant", "davo", "attend", "summit" e "davos" como os principais representantes do tópico, levando a identificar subjetivamente o tópico como "Fórum Econômico Mundial".

O tópico 6 trata de "tax", "rich", "billionaire", "wealth", "pay" e "oxfam" como construtoras do sentido majoritário do tópico. Por conta disso, o tópico será compreendido subjetivamente como "Riqueza e Taxação". O tópico 7 é aquele que apresenta a melhor distribuição entre as palavras, sendo a máxima não representando mais do que 12% do corpus. As palavras "export", "port", "black sea" e "ship" denotaram o aspecto do comércio internacional, enquanto "food", "grain" e "hunger" denotam o sentido de "Fome e Alimentação". Assim, o tópico 7 será chamado de "Comércio internacional, Alimentação e Fome".

Por conseguinte, é evidente que o tópico 2 trata dos aspectos da Guerra da Ucrânia, pois apresenta significativamente as palavras, "Ukranian", "Russian", "military", "attack", "force", "weapon", "soldier" e "sanction". Finalmente, é evidente que o tópico 1 trata dos aspectos da pandemia da covid-19, por contar com "vaccine", "health", "covid", "pfizer" e "treatment" como principais palavras representativas.

É sobre o viés das palavras mais frequentes pelos documentos mais representativos de cada tópico que se pôde denotar um conceito para cada tópico. A tabela 4 sintetiza o nome dado aos tópicos do modelo final com 9 tópicos.

Tabela 4 – Síntese do modelo 9

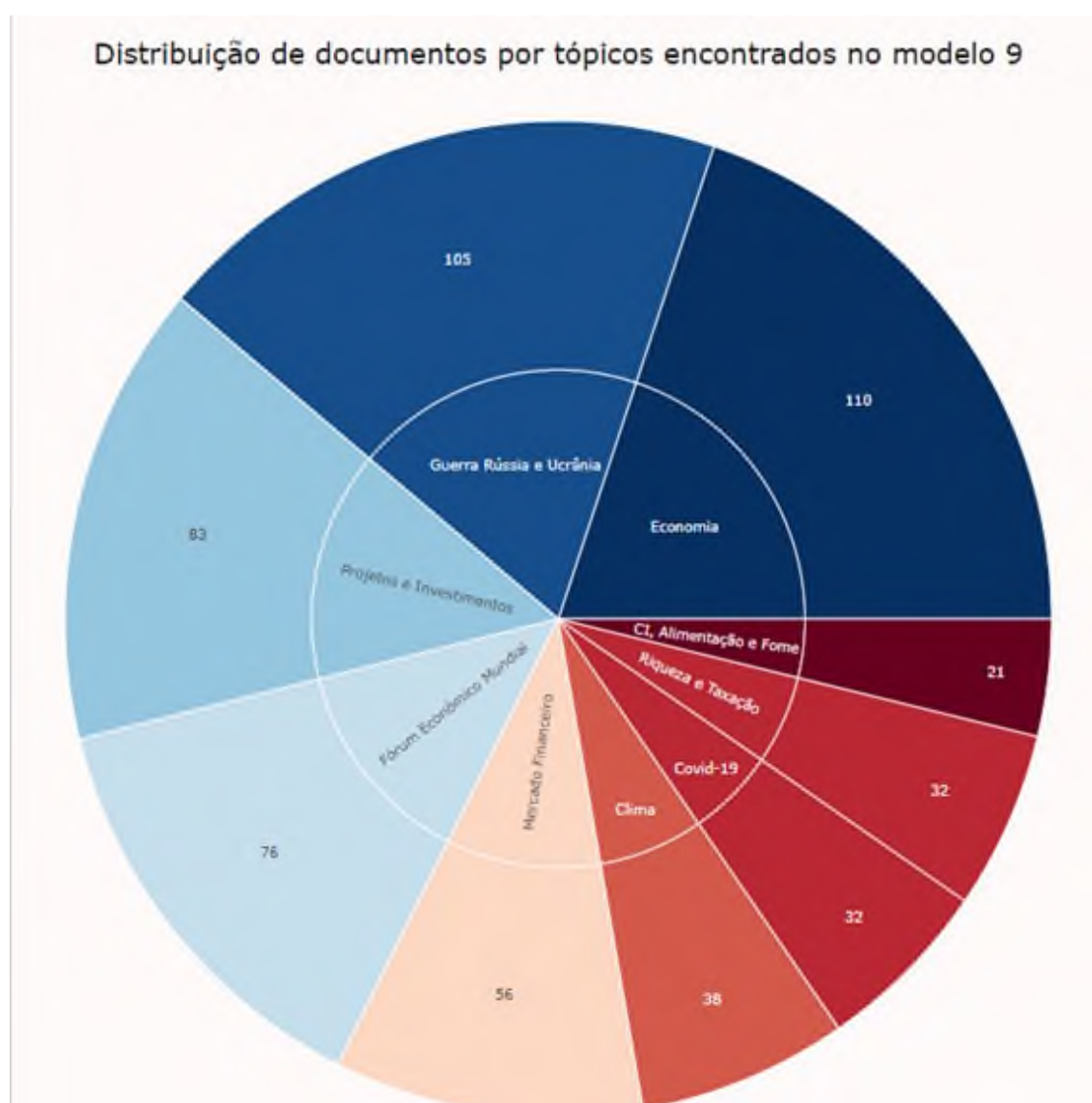
Tópico	Provável conceito
0	Economia
1	Covid-19
2	Guerra Rússia e Ucrânia
3	Projetos e Investimentos
4	Clima
5	Fórum Econômico Mundial
6	Riqueza e Taxação
7	Comércio Internacional, Alimentação e Fome
8	Mercado Financeiro

Fonte: elaboração própria com base nos dados obtidos do modelo 5

4.2 Resultados da modelagem de tópicos

Tomando como base o critério de escolha de seleção do número de tópicos e a semântica subjetiva utilizada para nomear cada um dos tópicos, pode-se observar a distribuição de documentos por tópicos na figura 11. Nela observa-se sequencialmente a distribuição geral de documentos por tópicos. Observa-se que Economia, Guerra da Rússia e Ucrânia, Projetos e Investimentos e Fórum Econômico Mundial são os principais temas prevalecentes. Em seguida, Clima; Covid-19; Riqueza e Taxação; e Comércio Internacional, Alimentação e Fome são aqueles que compõe marginalmente o corpus deste estudo.

Figura 11 – Distribuição absoluta de documentos por tópicos



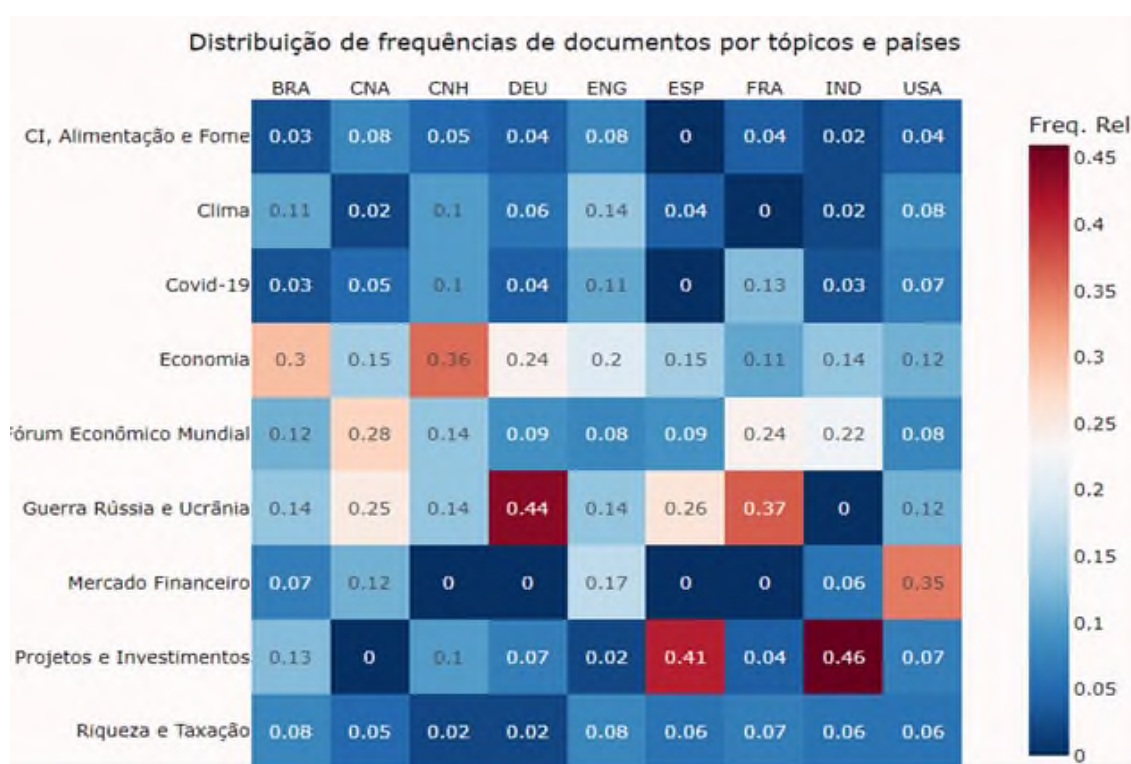
Fonte: elaboração própria com base nos dados obtidos da experimentação

Com base na figura 11, a figura 12 traz a distribuição relativa desses tópicos por países. Observa-se na figura que, relativamente, grupos de países apresentam

sobrevalência de interesses em tópicos diversos. Enquanto Espanha e Índia têm o tópico "Projetos e Investimentos", com respectivamente 0.41 e 0.43; Alemanha e França têm "Guerra Rússia e Ucrânia", com 0.44 e 0.37; China, Brasil e Inglaterra têm "Economia", com 0.3, 0.36 e 0.24; Canadá têm os aspectos gerais do "Fórum Econômico Mundial", com 0.28; e Estados Unidos têm Mercado Financeiro com 0.35.

Na figura, também é possível notar que a depender do país, o 1/3 dos tópicos mais prevaletentes é mais ou menos concentrado e superior a 0.5. Na sequência, classificam-se Espanha (0.82), Índia (0.82), Alemanha (0.77), França (0.74), Canadá (0.68), China (0.64), USA (0.59), Brasil (0.57) e Inglaterra (0.51).

Figura 12 – Distribuição relativa de tópicos por países



Fonte: elaboração própria com base nos dados obtidos da experimentação

Desses três tópicos, o segundo mais predominante relativamente à cada país é: "Guerra Rússia e Ucrânia" para Espanha, Canadá, China, Estados Unidos e Brasil; "Fórum Econômico Mundial" para China, França e Índia; "Economia" para Alemanha, e Estados Unidos; e, isoladamente, "Mercado Financeiro" para Inglaterra. Observa-se a repetição de China e Estados Unidos, pois ambos possuem a segunda maior distribuição relativa para cada dois tópicos.

O terceiro tópico mais predominante relativo à cada país também varia. A mídia da Espanha, Índia e Canadá se interessou mais por "Economia". Já na Inglaterra, Estados Unidos e China, a mídia teve mais interesse em "Clima". Por conseguinte, Brasil e China compartilham "Projetos e Investimentos" como tópico

terciário, enquanto a mesma China e França compartilham "Covid-19" e Estados Unidos e Alemanha "Fórum Econômico Mundial". Finalmente, Inglaterra apresentou interesse em "Guerra da Rússia e Ucrânia". Observa-se que China se repete três vezes e Inglaterra e Estados Unidos duas vezes. Isso acontece porque o terceiro tópico tem a mesma proporcionalidade entre três e dois tópicos, respectivamente.

Assim, observa-se que todos os países têm uma concentração relativa de pelo menos 50% do Corpus entre os três primeiros tópicos, mas esses tópicos se alteram nos diferentes níveis de importância relativa. No primeiro nível, por duas vezes um tópico predominou em dois países e uma vez para três, mostrando que o compartilhamento predominante de pelo menos um tópico recobre 77% dos países selecionados.

Ademais, observa-se que "Economia" prevaleceu entre os três tópicos com maior cobertura somando 8 vezes, sendo três vezes como primário, cinco como secundário e uma como terciário. Em seguida, Guerra da Rússia e Ucrânia somam cada uma 6 vezes, sendo o primeiro ocorrendo duas vezes como primária, cinco como secundária e uma como terciária; e o segundo, uma vez como primário, três como secundário e dois como terciário. Por fim, Clima e Covid-19 só apareceram como temas terciários, cada um com frequência de duas vezes.

4.3 Resultados da modelagem de emoções

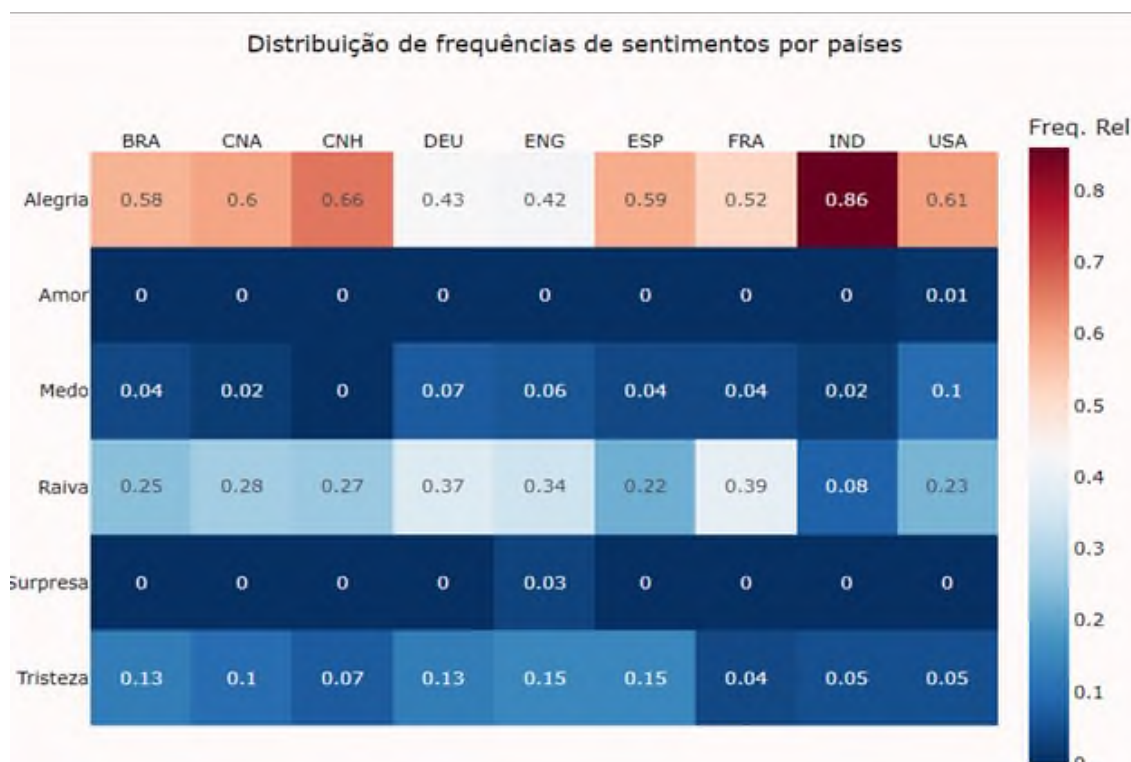
Dotado de modelo pré-validado em contexto diverso e abrangente, os resultados obtidos para o algoritmo de análise de emoções pode ser esboçado na figura 13. Como aponta a figura, "Alegria" é o sentimento predominante para todos os países proporcionalmente. Em seguida, "Raiva" aparece como a segunda emoção mais predominante entre os países. Por conseguinte, "Tristeza" e "Medo" são as terceiras emoções mais predominantes relativas à cada país.

Na figura, também é possível constatar que todos os países evocam o mesmo padrão consecutivo de "Alegria", "Raiva" e "Tristeza", com exceção dos Estados Unidos, que evoca "Alegria", "Raiva" e "Medo"; e França que apresentam "Alegria", "Raiva", "Tristeza" e "Medo".

Em seguida, a figura 14 traz a distribuição relativa a cada país das emoções por tópicos. Alegria permanece o sentimento predominante na grande maioria dos tópicos, com exceção de Guerra da Rússia e da Ucrânia e Comércio Internacional, Alimentação e Fome, o qual "Raiva" foi o sentimento predominante.

A combinação das três emoções predominantes segue apenas três padrões compartilhados: "Alegria", "Raiva" e "Tristeza" são, nessa sequência, as três emoções

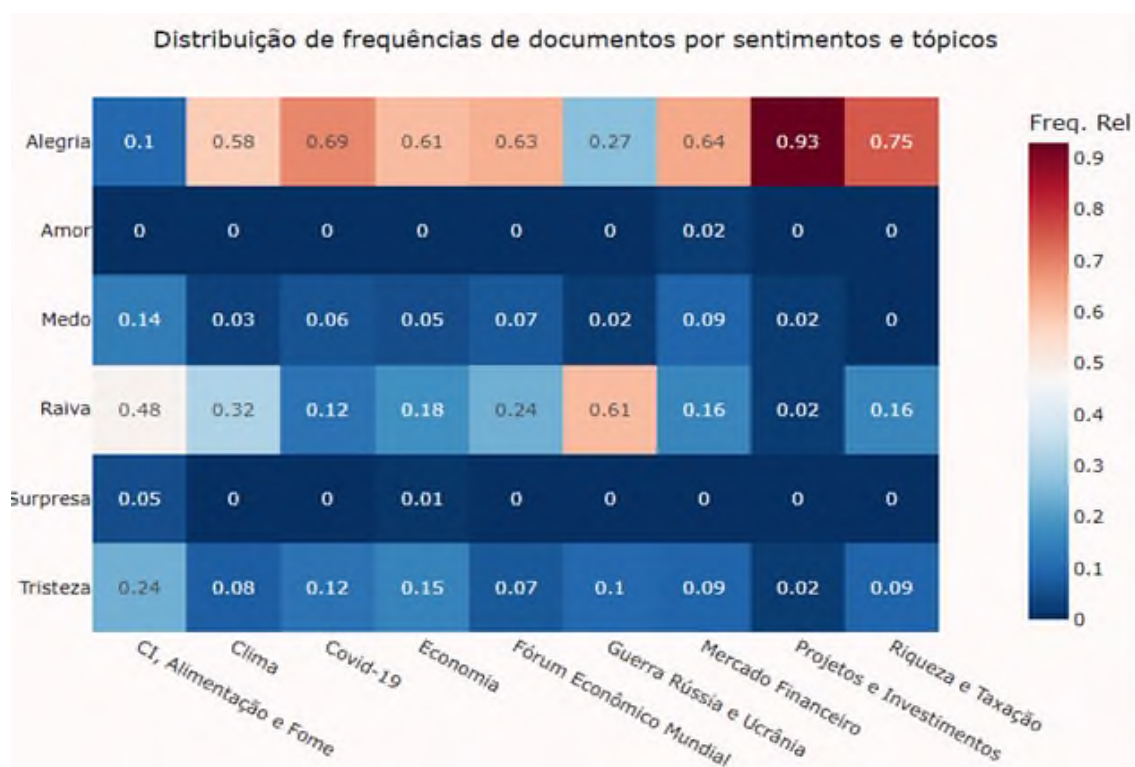
Figura 13 – Distribuição relativa de emoções por países



Fonte: elaboração própria com base nos dados obtidos da experimentação

predominantes para "Clima", "Covid-19", "Economia" e "Riqueza e Taxação"; "Raiva", "Tristeza" e "Medo" são predominantes para os temas de "Guerra Rússia e Ucrânia" e "Comércio Internacional, Alimentação e Fome"; e "Alegria" e "Raiva", com "Tristeza" e "Medo" relativamente iguais, para "Fórum Econômico Mundial", "Mercado Financeiro" e "Projetos e Investimentos". Finalmente, observa-se que "Amor" e "Surpresa" sempre foram emoções relativamente marginais.

Figura 14 – Distribuição relativa de emoções por tópicos



Fonte: elaboração própria com base nos dados obtidos da experimentação

5 Conclusão

Como exposto na segunda seção deste trabalho, o Fórum Econômico Mundial é a organização internacional voltada para a parceria público-privada, que conta com stakeholders diversos, capacidade de ação e objetivos na construção da governança global para além do escopo estatocêntrico. A pluralidade de atores, o desenvolvimento histórico-institucional e o engajamento são marcas fundamentais para que a instituição tenha se mantido viva desde 1970. Essas características foram fundamentais para criação das comunidades epistêmicas que compõem o FEM.

Nesse contexto, apesar de pouco estudada e teorizada, a mídia possui papel essencial na co-produção do espaço epistêmico e na difusão global de ideias, fatos e acontecimentos discutidos dentro do Fórum. Assim, através de suas capacidades de poder e influência, a mídia é capaz de engajar e formar parcialmente a opinião pública através da cobertura midiática. Outro aspecto fundamental é sua capacidade de gerar prestígio aos atores envolvidos e co-legitimar apreciações.

Foi buscando entender como a mídia atuou no evento de Davos em 2022 que este trabalho se iniciou. Através de modelos de Machine Learning e Deep Learning para modelagem de tópicos e análise de emoções se procurou explorar de que forma se deu a difusão global de notícias em diferentes países, medido por diferentes fontes.

Após o procedimento comparativo com a métrica de coerência, pôde-se identificar que LDA performou de forma bastante superior aos modelos de Deep Learning. Em seguida, selecionando os menores mínimos de RTPC pôde se observar 3 modelos candidatos à número de tópicos. O primeiro modelo, com 5 tópicos, mostrou-se restritivo por combinar temas que à priori não pareciam ter uma ligação direta. Por sua vez, o terceiro modelo se mostrou demasiadamente específico, repetindo por vezes o mesmo tópico com poucas palavras novas e criando tópicos com baixa probabilidade condicional.

Posto isso, o modelo selecionado foi o de 9 tópicos, pois era aquele dentre os três que pôde separar tópicos sem ser muito geral, nem muito específico. Os tópicos que foram descobertos passaram por uma análise subjetiva para receberem um nome específico baseado na frequência das 10 palavras mais significativas dos 20 documentos mais significativos. Os nomes dos tópicos, sequenciados pelo número de documentos foram: "Economia", "Guerra da Rússia e Ucrânia", "Projetos e Investimentos", "Fórum Econômico Mundial", "Mercado Financeiro", "Clima", "Covid-19", "Riqueza e Taxação" e "Comércio Internacional, Alimentação e Fome".

Quando olhado pelo eixo da distribuição relativa à cada país de frequências de tópicos predominantes, observou-se que mais da metade das notícias difundidas para cada país podem ser divididas em até três tópicos. Desses, "Economia" predominou para os casos do Brasil, China e Inglaterra; "Projetos e Investimentos" para Espanha e Índia; "Guerra da Rússia e Ucrânia" para Alemanha e França. Canadá e Estados Unidos tiveram respectivamente "Fórum Econômico Mundial" e "Mercado Financeiro".

Em seguida, "Guerra da Rússia e Ucrânia" sempre ocupou a maior parte das vezes a posição secundária em importância. Em especial ao tema de "Guerra da Rússia e Ucrânia" é interessante notar que países mais distantes do epicentro do conflito como China, Estados Unidos e Brasil esboçam interesse secundário no tópico. "Fórum Econômico Mundial" também segue o mesmo comportamento como o tópico mais secundariamente mais frequente. Finalmente, pôde se constatar que "Clima" e "Covid-19" sempre foram temas terciários, somando 4 aparições.

Nesse ponto, é interessante notar que em 2021 o Fórum diagnosticou a Covid-19 como as bases para "O grande recomeço", mas esse chegou apenas a ser o terceiro interesse mais importante difundido.

Quando generalizada a visão para o top-3 tópicos, observa-se que "Economia" ocupou 8 posições, sendo três vezes o principal tópico, duas vezes o tópico secundário e três vezes o terciário. Isso indica uma característica fundamental do interesse de cobertura compartilhada pela maioria da mídia dos países analisados.

Quando olhamos pelo eixo das emoções, o sentimento geral para esses tópicos foi de "Alegria", com exceção de "Guerra da Rússia e Ucrânia" e "Comércio Internacional, Alimentação e Fome", nos quais "Raiva" foi a emoção predominante. Para entender se de fato esses dois tópicos compartilham relações, seria preciso uma futura análise qualitativa mais aprofundada para cada notícia clusterizada.

Quanto as limitações da pesquisa, observa-se em primeiro lugar o problema do próprio campo de "modelagem de tópicos" em criar uma métrica que possa sintetizar o modelo. Muitos avanços foram alcançados, mas ainda carece um framework que consiga unir os avanços obtidos e dar capacidade de meios de avaliação que são desenvolvidos no "estado da arte". No pacote OCTIS do Python, Terragni et al (2021) fez um brilhante trabalho de trazer ambos modelos mais recentes e métricas contemporâneas, mas ainda falta um pouco de acessibilidade para o usuário menos experiente. Por outro lado, o pacote Tomotopy conseguiu trazer mais acessibilidade e possibilidades de insights detalhados sobre tópicos e documentos, mas está apenas limitado ao contexto de Machine Learning.

Em segundo lugar, recursos computacionais devem ser levados em conta, apesar de técnicas mais aprofundadas de redução de dimensionalidade que em

contexto de menos dados podem entregar o mesmo resultado.

Finalmente, o processo de subjetivação deve ser feito com cautela, pois ainda que subjetivo, deva seguir algum critério que possa ser replicado, contestado e melhorado. Nesse contexto, seria essencial a co-validação entre pares, a qual pode ser realizada em desenvolvimentos futuros.

Ao fim, reconhece-se que o método de modelagem de tópico e de análise de emoções para explorar a cobertura midiática no contexto de Organizações Internacionais parece promissor. Uma possível continuidade da pesquisa seria sua extensão temporal para saber o "tempo de vida" de cada tópico. Além disso, a replicação do método para outros contextos é realizável, desde que sejam baseados em tipos de dados textuais não estruturados. Essa área de modelagem de tópicos permanece ativa com desenvolvimentos no "estado da arte" sendo disponibilizado anualmente à toda comunidade.

Referências

- ALETRAS, Nikolaos; STEVENSON, Mark. Evaluating topic coherence using distributional semantics. In: Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers. 2013. p. 13-22.
- ALFIERI, Luca; GABRIELIAN, Diana. The Communication Reaction Function of the European Central Bank. an Analysis using Topic Modelling. Bank of Estonia Working Paper 8/2021, Tallinn, 2021.
- ANAYA, Leticia H. Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers. ProQuest LLC. 789 East Eisenhower Parkway, PO Box 1346, Ann Arbor, MI 48106, 2011.
- AVANT, Deborah D.; FINNEMORE, Martha; SELL, Susan K. (Ed.). Who governs the globe?. Cambridge University Press, 2010.
- BARKIN, Noah; PIPER, Elizabeth. In Davos, Xi makes case for Chinese leadership role. Reuters, [S. l.], p. 0-10, 17 jan. 2017.
- BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent dirichlet allocation. Journal of machine Learning research, v. 3, n. Jan, p. 993-1022, 2003.
- BROWNLEE, Jason. A GENTLE Introduction to Dropout for Regularizing Deep Neural Networks. Machine Learning Mastery, [S. l.], p. 0-10, 3 dez. 2018. Disponível em: <https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/>. Acesso em: 14 nov. 2022./S1519-38292010000100001.
- CABRAL FILHO, J E. Revista Brasileira de Saúde Materno Infantil: dez anos. Revista Brasileira de Saúde Materno Infantil. [S. l.]: FapUNIFESP (SciELO), mar. 2010. DOI 10.1590/s1519-38292010000100001. Disponível em: <http://dx.doi.org/10.1590>
- CAMPAGNOLO, João Marcos; DUARTE, Denio; DAL BIANCO, Guilherme. Topic Coherence Metrics: How Sensitive Are They?. Journal of Information and Data Management, v. 13, n. 4, 2022.
- CHANG, Jonathan et al. Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems, v. 22, 2009.
- CRAWLEY, Sam; COFFÉ, Hilde; CHAPMAN, Ralph. Public opinion on climate change: Belief and concern, issue salience and support for government action. The British Journal of Politics and International Relations, v. 22, n. 1, p. 102-121, 2020.

DAVOS Annual Meeting 2005 - Luiz Inacio Lula da Silva. [S. l.: s. n.], 2005. Disponível em: <https://www.youtube.com/watch?v=XyQldu-Pst0>. Acesso em: 24 out. 2022.

DAVOS Annual Meeting 2009 - Vladimir Putin. [S. l.: s. n.], 2009. Disponível em: <https://www.youtube.com/watch?v=iy-el0jO1mo>. Acesso em: 24 out. 2022.

DEVLIN, Jacob et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

EDISON, Hali; CARCEL, Hector. Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts. *Applied Economics Letters*, v. 28, n. 1, p. 38-42, 2021.

FAINBERG, Alisa. Spread the word: Russia social media on the service of jihad. The International Institute for Counter-Terrorism (ICT), 2017.

FRIESEN, Elizabeth. The World Economic Forum and Transnational Networking. Emerald Group Publishing, 2020.

GILBOA, Eytan. Media diplomacy: Conceptual divergence and applications. *Harvard International Journal of Press/Politics*, v. 3, n. 3, p. 56-75, 1998.

GILBOA, Eytan. Global communication and foreign policy. *Journal of communication*, v. 52, n. 4, p. 731-748, 2002.

GLOBO, Rede. Em Davos, Greta Thunberg pede que líderes mundiais ouçam jovens ativistas. *Globo*, [S. l.], p. 0-10, 21 jan. 2020. Disponível em: <https://g1.globo.com/natureza/noticia/2020/01/21/em-davos-greta-thunberg-pede-que-lideres-mu>. Acesso em: 25 out. 2022.

GUTERRES, António. António Guterres: Read the UN Secretary-General's Davos speech in full. *World economic forum*, [S. l.], p. 0-10, 24 jan. 2019. Disponível em: <https://www.weforum.org/agenda/2019/01/these-are-the-global-priorities-and-risks-for-the-future->. Acesso em: 24 out. 2022.

HENING, Phillip. Probabilistic ML — Lecture 20 — Latent Dirichlet Allocation. YouTube, 2021, Disponível em: <https://www.youtube.com/watch?v=z2q7LhsnWNg>. Acesso em: 6 set. 2022.

HIGGOTT, Richard A.; UNDERHILL, Geoffrey RD; BIELER, Andreas (Ed.). *Non-state actors and authority in the global system*. 2000.

JIANG, Ye et al. Comparing Attitudes to Climate Change in the Media using sentiment analysis based on Latent Dirichlet Allocation. In: *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*. 2017. p. 25-30.

KFIR, Isaac. Social identity group and human (in) security: The case of Islamic State in Iraq and the Levant (ISIL). *Studies in Conflict Terrorism*, v. 38, n. 4, p. 233-252, 2015.

KHANNA, Parag. *The World Economic Forum: An anatomy of multi-stakeholder global policy-making*. London School of Economics and Political Science (United Kingdom), 2010.

KHERWA, Pooja; BANSAL, Poonam. Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, v. 7, n. 24, 2019.

KORENČIĆ, Damir; RISTOV, Strahil; ŠNAJDER, Jan. Document-based topic coherence measures for news media text. *Expert systems with Applications*, v. 114, p. 357-373, 2018.

KORSTANJE, J. The F1 score. 2021. Disponível em: <<https://towardsdatascience.com/the-f1-score-bec2bbc38aa6>>. Acesso em: 15 nov. 2022.

KRASNER, Stephen D. (Ed.). *International regimes*. Cornell University Press, 1983.

LAN, Zhenzhong et al. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

LAU, Jey Han et al. Best topic word selection for topic labelling. In: *Coling 2010: Posters*. 2010. p. 605-613.

LINDER, Stephen H.; ROSENAU, Pauline Vaillancourt. *Mapping the terrain of the public-private policy partnership*. 2000.

LIU, Yinhan et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 2019. DOI: 10.48550/ARXIV.1907.11692. Disponível em: <https://arxiv.org/abs/1907.11692>.

MARKOVITZ, Dale. EXPLICAÇÃO sobre transformadores: entenda o modelo por trás de GPT, BERT, e T5. [S. l.]: Google Cloud Tech, 2022. Disponível em: <https://www.youtube.com/watch?v=SZorAJ4I-sA>. Acesso em: 15 nov. 2022.

MARSHALL, Andrew. *World Economic Forum: a history and analysis*. Transnational Institute, [S. l.], p. 0-10, 1 jan. 2015. Disponível em: <https://www.tni.org/es/node/84>. Acesso em: 24 out. 2022.

MAZIS, Panagiotis; TSEKREKOS, Andrianos. Latent semantic analysis of the FOMC statements. *Review of Accounting and Finance*, 2017.

MEI, Qiaozhu; SHEN, Xuehua; ZHAI, ChengXiang. Automatic labeling of multinomial topic models. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. 2007. p. 490-499.

MONK, Adrian. World Economic Forum 2021 – Highlights: Three key takeaways from Davos 2021. In: PACIFIC BASIN ECONOMIC COUNCIL. World Economic Forum 2021 – Highlights: : Three key takeaways from Davos 2021. [S. l.], 11 jan. 2022. Disponível em: <https://www.pbec.org/world-economic-forum-2021-highlights-three-key-takeaways-from-davos-2021>. Acesso em: 25 out. 2022.

MONTAVON, Grégoire; SAMEK, Wojciech; MÜLLER, Klaus-Robert. Methods for interpreting and understanding deep neural networks. Digital signal processing, v. 73, p. 1-15, 2018.

MOSTAFA, Bossy et al. Machine and Deep Learning Approaches in Genome: Review Article. Alfarama Journal of Basic Applied Sciences 0, 0 (August 2020), 0–0. DOI: <https://doi.org/10.21608/ajbas>, v. 1023, 2020.

NYE, Joseph S.; KEOHANE, Robert O. Transnational relations and world politics: An introduction. International organization, v. 25, n. 3, p. 329-349, 1971.

NYE JR, Joseph S.; OWENS, William A. America's information edge. Foreign Aff., v. 75, p. 20, 1996.

O'GORMAN, Kevin; MACINTOSH, Robert. Research methods for business and management: A guide to writing your dissertation. Goodfellow Publishers Ltd, 2015.

PARKER, Ceri. The World Economic Forum at 50: A timeline of highlights from Davos and beyond. World economic forum , [S. l.], p. 0-10, 20 out. 2019. Disponível em: <https://www.weforum.org/agenda/2019/12/world-economic-forum-davos-at-50-history-a-timeline-of-highlights>. Acesso em: 24 out. 2022.

PATHAK, Neetu; HENRY, Michael J.; VOLKOVA, Svitlana. Understanding Social Media's Take on Climate Change through Large-Scale Analysis of Targeted Opinions and Emotions. In: 2017 AAAI Spring Symposium Series. 2017.

PEDRO, João. Understanding Topic Coherence Measures: Learn the details behind this everyday tool from the Topic Modeling toolbox. Towards Data Science, [S. l.], p. 0-10, 3 jan. 2010. Disponível em: <https://towardsdatascience.com/understanding-topic-coherence-measures-4aa41339634c>. Acesso em: 14 nov. 2022.

PIGMAN, Geoffrey Allen. The World Economic Forum: A Multi-Stakeholder Approach to Global Governance. [S. l.: s. n.], 2007.

PRABHU, P. Understanding neurons in deep learning. 2020. Disponível em: <<https://medium.com/analytics-vidhya/understanding-neurons-in-deep-learning-8308ce91271f>>. Acesso em: 15 nov. 2022.

RAJPURKAR, P. et al. SQuAD: 100,000+ Questions for Machine Comprehension of TextarXiv, , 2016. . Disponível em: <<https://arxiv.org/abs/1606.05250>>.

RAJPURKAR, P.; JIA, R.; LIANG, P. Know What You Don't Know: Unanswerable Questions for SQuADarXiv, , 2018. . Disponível em: <<https://arxiv.org/abs/1806.03822>>.

ROBINSON, Olga et al. 'Great Reset': como plano econômico virou teoria da conspiração global. BBC News, [S. l.], p. 0-10, 30 jun. 2021. Disponível em: <https://www.bbc.com/portuguese/internacional-57666500>. Acesso em: 25 out. 2022.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning representations by back-propagating errors. nature, v. 323, n. 6088, p. 533-536, 1986.

SANH, Victor et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

SNYDER, Justin et al. Topic models and metadata for visualizing text corpora. In: Proceedings of the 2013 NAACL HLT Demonstration Session. 2013. p. 5-9.

SYED, Shaheen; SPRUIT, Marco. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In: 2017 IEEE International conference on data science and advanced analytics (DSAA). IEEE, 2017. p. 165-174.

TERRAGNI, Silvia; FERSINI, Elisabetta. OCTIS 2.0: Optimizing and Comparing Topic Models in Italian Is Even Simpler!. In: CLiC-it. 2021.

VASILIEV, Aleksei M. et al. Instruments of the Islamic State Soft Power: Typology and Performance Evaluation. Asia and Africa Today, n. 12, p. 3-10, 2018.

VASWANI, A. et al. Attention Is All You NeedarXiv, , 2017. . Disponível em: <<https://arxiv.org/abs/1706.03762>>.

WANG, W.; YAN, M.; WU, C. Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question AnsweringProceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Em: PROCEEDINGS OF THE 56TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (VOLUME 1: LONG PAPERS). Association for Computational Linguistics, , 2018. . Disponível em: <<http://dx.doi.org/10.18653/v1/P18-1158>>.

WANG, Alex et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461, 2018.

WANTA, Wayne; GOLAN, Guy; LEE, Cheolhan. Agenda setting and international news: Media influence on public perceptions of foreign nations. *Journalism Mass Communication Quarterly*, v. 81, n. 2, p. 364-377, 2004.

WILSON, Andrew; CHEW, Peter A. Term weighting schemes for latent dirichlet allocation. In: *human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics*. 2010. p. 465-473.

THE DAVOS Agenda. In: SUÍÇA (Davos). World Economic Forum. *World economic forum* . [S. l.], 29 jan. 2021. Disponível em: <https://www.weforum.org/events/the-davos-agenda-2021>.

WORLD ECONOMIC FORUM. Slideshare. In: *World economic forum Annual Report : 2001/2002*. [S. l.], 29 jan. 2009. Disponível em: <https://www.slideshare.net/WorldEconomicForumDavos/world-economic-forum-annual-report-2001-2002>. Acesso em: 24 out. 2022.

WORLD ECONOMIC FORUM , Davos. *Annual Repport 2016-2017*. [S. l.: s. n.], 2017.

CHARLTON, Emma. Fake-news-what-it-is-and-how-to-spot-it. *World Economic Forum*, [S. l.], p. 0-10, 6 mar. 2019. Disponível em: <https://www.weforum.org/agenda/2019/03/fake-news-what-it-is->

WORLD Economic Forum 2020 | Highlights from Davos. Direção: IESE Business School. [S. l.: s. n.], 2020. Disponível em: <https://www.youtube.com/watch?v=KTJa03LphtY>. Acesso em: 25 out. 2022.

World Economic Forum. *A Platform for Impact*. [S. l.: s. n.], 2022.

ZELLERS, R. et al. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Em: *PROCEEDINGS OF THE 2018 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING*. Association for Computational Linguistics, , 2018. . Disponível em: <<http://dx.doi.org/10.18653/v1/D18-1009>>.

ZIZEK, Slavoj. *Welcome to the desert of the real: Five essays on September 11 and related dates*. Verso Books, 2013.