REVISTA FATEC SEBRAE EM DEBATE:
Gestão, Tecnologias e Negócios
Volume 08 | Número 15 | JUL. – DEZ. 2021 | ISSN: 2358-9817

O IMPACTO DA PANDEMIA NOS MICRO E PEQUENOS EMPRESÁRIOS: APLICAÇÃO DE MODELAGEM DE TÓPICOS EM COMENTÁRIOS NO INSTAGRAM

Júlio Boaro
Mauro Cabral
Orientador: Prof. Alexander Homenko

Resumo

A presente pesquisa exploratória consiste na aplicação de técnica de modelagem de tópicos, denominada *Latent Dirichlet Allocation* ou LDA, em comentários publicados na plataforma de mídia social Instagram em posts de perfis dos principais veículos editoriais de notícias que publicam sobre empreendedorismo, negócios e temas relacionados. Foram coletados posts e comentários de um total de 9 perfis ao longo de 17 meses de pandemia, que foram tratados para análise. A aplicação da técnica de modelagem de tópicos permitiu identificar diferentes temas abordados facilitando a análise exploratória do impacto que a pandemia causou e que foi comentado pelos micro e pequenos empreendedores no Instagram. A técnica se mostrou eficaz e adequada para uma análise exploratória em um contexto de big data, permitindo a seleção de uma amostra de comentários para leitura analítica entre dezenas de milhares, agrupada por tópicos, e sem os vieses de leitura da realidade se considerasse apenas os conteúdos que nos chegam mediados pelos algoritmos das plataformas, como os conteúdos virais.

Palavras-chave: Modelagem de Tópicos, Análise de Redes Sociais, Negócios, Empreendedorismo, Pandemia.

FATEC Sebrae – Faculdade de Tecnologia Sebrae - CEETEPS – Centro Estadual de Educação Tecnológica Paula Souza – São Paulo, Brasil.

REVISTA FATEC SEBRAE EM DEBATE: gestão, tecnologias e negócios

Editor Geral
Prof. Dr. Mário Pereira Roque Filho

Organização e Gestão
Prof. Ms. Clayton Pedro Capellari

Correspondência
Alameda Nothmann, nº 598 Campos Elíseos, CEP 01216-000 São Paulo – SP, Brasil.
+55 (11) 3224.0889 ramal: 218
E-mail: f272dir@cps.sp.gov.br

Abstract

This exploratory research consists of applying a topic modeling technique, called Latent Dirichlet Allocation or LDA, in comments published on the social media platform Instagram in profile posts of the main editorial news vehicles that publish about entrepreneurship, business and related topics. Posts and comments were collected from a total of 9 profiles over the 17 months of the pandemic, which were processed for analysis. The application of the topic modeling technique allowed us to identify different topics covered, facilitating the exploratory analysis of the impact that the pandemic caused and that was commented on by micro and small entrepreneurs on Instagram. The technique proved to be effective and adequate for an exploratory analysis in a big data context, allowing the selection of a sample of comments for analytical reading among tens of thousands, grouped by topics, and without the biases of reading reality if only the contents that reach us mediated by platform algorithms, such as viral contents.

Keywords: Topic Modeling, Social Networking Analysis, Business, Entrepreneurship, Pandemic.

Introdução

Introdução A pandemia de Covid-19 abalou o mundo dos negócios e as relações de trabalho. Nesta pesquisa buscamos identificar os impactos da pandemia nos gestores, empresários e empreendedores de modo indireto, isto é, através de relatos espontâneos deixados em comentários nos posts de perfis no Instagram que publicam notícias e informações na área de negócios e empreendedorismo. Considerando que as pessoas que seguem tais perfis são micro e pequenos empreendedores, seria possível mapear e identificar os principais assuntos comentados por estas pessoas nos posts, no período e no contexto da pandemia? Como, a partir de dezenas de milhares de comentários, poderíamos obter uma síntese que permitisse uma leitura geral do impacto da pandemia para este público?

Para responder a estas perguntas e para a execução de uma classificação eficiente, atualizada com o “estado da técnica”, aplicamos uma metodologia robusta que utiliza todo o corpus dos comentários, considerando cada termo e suas co-ocorrências.

Objetivo

Aplicação de uma técnica de modelagem de tópicos que permita melhor compreender os principais assuntos que foram comentados pelos empreendedores durante a pandemia.

Justificativa

As redes sociais são um espaço valioso de expressão que deve ser explorado para melhor compreender as opiniões e comportamentos que geram impactos nos negócios.

Fundamentação teórica

Modelagem de tópicos

Faleiros (2016) descreve tópicos como:

Os tópicos são estruturas com valor semântico e que, no contexto de mineração de texto, formam grupos de palavras que frequentemente ocorrem juntas. Esses grupos de palavras, quando analisados, dão indícios a um tema ou assunto que ocorre em um subconjunto de documentos. A expressão tópico é usada levando-se em conta que o assunto tratado em uma coleção de documentos é extraído automaticamente, ou seja, tópico é definido como um conjunto de palavras que frequentemente ocorrem em documentos semanticamente relacionados.

Então, a modelagem de tópicos relaciona-se ao processo de geração destes tópicos a partir de um grande conjunto de documentos, que no caso desta pesquisa são comentários deixados em rede social.

LDA - “Latent Dirichlet Allocation”

Para esta pesquisa aplicamos um tipo de técnica de modelagem de tópicos denominada “Latent Dirichlet Allocation”, ou simplesmente LDA, que foi desenvolvida por Blei (2003), sendo há muito testada e aprovada, ao ponto de ser uma das mais técnicas mais populares de modelagem de tópicos entre os profissionais da área de Ciência de Dados.

A técnica de modelagem de tópicos com LDA trata cada documento, no nosso caso comentários de Instagram, como formado por uma mistura de tópicos, e cada tópico como formado por uma mistura de termos.

Faleiros (2016) definiu LDA como:

> O LDA é um modelo probabilístico generativo para coleções de dados discretos como corpus de documentos (BLEI; NG; JORDAN, 2003). Um modelo generativo é aquele que aleatoriamente gera os dados a partir das variáveis latentes. Assim, o LDA não é um algoritmo com descrições sequenciais de instruções para encontrar tópicos dada uma coleção de documentos. O LDA é um modelo probabilístico no qual é descrito como os documentos são gerados. Nesse modelo, as variáveis observáveis são os termos de cada documento e as variáveis não observáveis são as distribuições de tópicos. Os parâmetros das distribuições de tópicos, conhecidos como hiper-parâmetros, são dados a priori no modelo.

De maneira simplificada a modelagem de tópicos por LDA, é uma técnica de classificação de texto. Esta técnica consiste em analisar a estrutura dos textos, chamados de "documentos".

Para isso, os textos são agrupados em um único conjunto que será o objeto de análise para a técnica LDA. Este conjunto único é denominado “corpus”. O analista define a priori a quantidade de tópicos desejados, e pode iterativamente testar diferentes quantidades de tópicos permitindo avaliar a qualidade do resultado do modelo com base no conhecimento do analista sobre o assunto pesquisado.

A saída do modelo LDA são 2 resultados: (1) uma lista de probabilidades de cada documento estar associado a cada um dos tópicos, chamado “gamma” e (2) uma lista de probabilidades de cada termo do texto estar associado a cada um dos tópicos, o que considera todos os termos do texto, chamado “beta”.

Uma vantagem da técnica LDA é que um termo pode pertencer a vários tópicos e um documento, no nosso caso post, igualmente pode pertencer a vários tópicos, o que torna a modelagem mais suave e robusta. Em outras palavras, não é porque um documento tem similaridade com um tópico que também não pode ter com outro, essas associações são medida em probabilidades. O mesmo vale para os termos. Um termo

pode estar fortemente associado a um tópico, mas também pode estar fortemente associado a outro, o que não é desejável, mas pode ser observado em alguns casos.

Como LDA trabalha com probabilidades, então, o que difere um tópico do outro são os valores de probabilidades de ocorrência de cada um dos termos em cada tópico e, analogamente, os valores de probabilidades que os documentos recebem em estar associados a cada um dos tópicos.

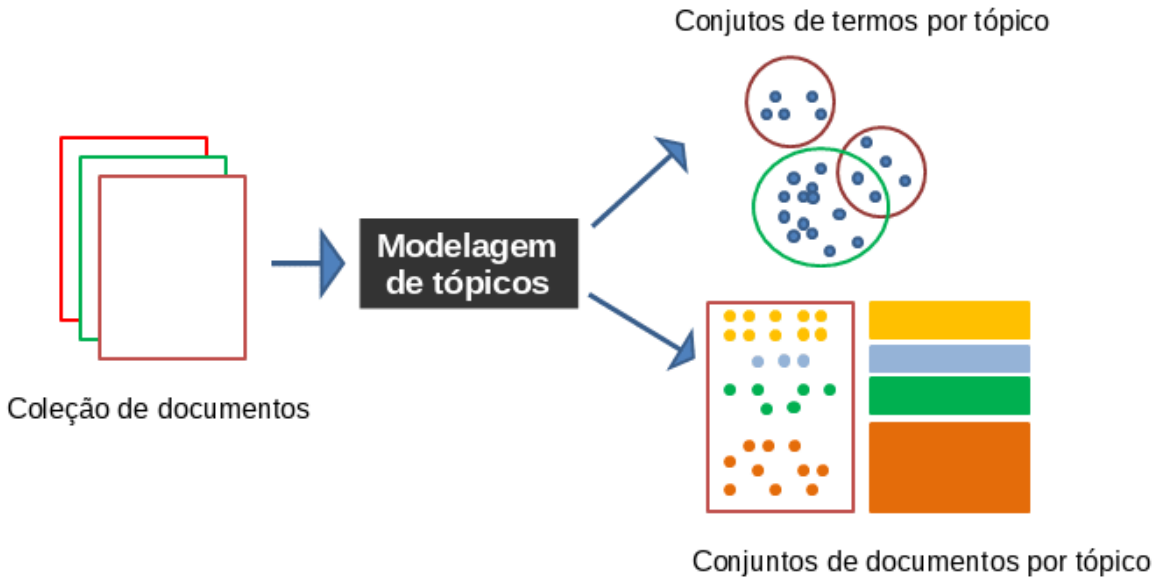

Figura 1 - Esquema de modelagem de tópicos com Latent Dirichlet Allocation (LDA). Adaptado de https://thinkinfi.com/latent-dirichlet-allocation-for-beginners-a-high-level-overview

Materiais e métodos (coleta e processamento dos dados)

Primeiramente, elencamos 9 dos principais perfis no Instagram de projetos editoriais na área de negócios e empreendedorismo. São eles:

- @administradores
- @endeavorbrasil
- @epocanegocios
- @estadaopme

- @revistapegn
- @sebrae
- @sebraesp
- @startseoficial
- @vocesa

O perfil @hotmart fazia parte da seleção inicial mas foi retirado porque misturava português, inglês e espanhol no mesmo post, o que atrapalhava a etapa de preparação e limpeza dos dados.

Determinamos um período de análise iniciando em 1º de março de 2020 até 31 de julho de 2021, considerando que o 1º caso oficial de Covid-19 no Brasil é datado de 26 de fevereiro, já no final do mês, e a 1ª morte por Covid-19 data de 12 de março.

A ferramenta de coletada utilizada é o script Instaloader, software de código aberto desenvolvido em Python e disponível no Github.

Foi executada a coleta de posts dos perfis no período selecionado, utilizando script de raspagem de dados. Foram coletados 7.623 posts.

Para análise e processamento dos dados foi utilizada a linguagem R e o ambiente de desenvolvimento RStudio. Os posts foram filtrados de modo a delimitar o objeto de estudo em posts que contivessem a pandemia como um assunto diretamente mencionado. Foram filtrados apenas os posts com menção expressa e textual dos termos: “pandemia”, “coronavirus”, “covid”, o que inclui “covid-19”. Restaram então 1.630 posts, conforme a tabela:

Tabela 1 - Quantidade de posts coletados e selecionados por perfil

Perfil	**Posts**	**Posts pandemia**	**Posts pandemia (%)**
sebrae	2.009	570	28,4%

vocesa	1.196	270	22,6%
revistapegn	672	254	37,8%
epocanegocios	697	177	25,4%
administradores	1.304	150	11,5%
sebraesp	1.039	118	11,4%
startseoficial	374	46	12,3%
estadaopme	152	37	24,3%
endeavorbrasil	210	8	3,8%
Total	**7.653**	**1.630**	**21,2%**

Na etapa anterior, além dos posts, foram também coletados os comentários publicados nesses posts. O total de comentários coletados foi de 420.792. Considerando apenas os comentários em posts com menções à pandemia, que é o objeto principal desta análise, o total de comentários foi de 39.958, conforme a tabela:

Tabela 2 - Quantidade de comentários selecionados para análise por perfil

Perfil	**Comentários em posts pandemia**	**Comentários em posts pandemia (% da coluna)**
sebrae	11.541	28,9%
vocesa	10.260	25,7%
administradores	7.886	19,7%
revistapegn	4.808	12,0%

epocanegocios	3.139	7,9%
sebraesp	1.159	2,9%
startseoficial	1.047	2,6%
endeavorbrasil	63	0,2%
estadaopme	55	0,1%
Total	**39.958**	**100,0%**

A quantidade total mensal de comentários nos posts associados à pandemia pode ser observada no gráfico:

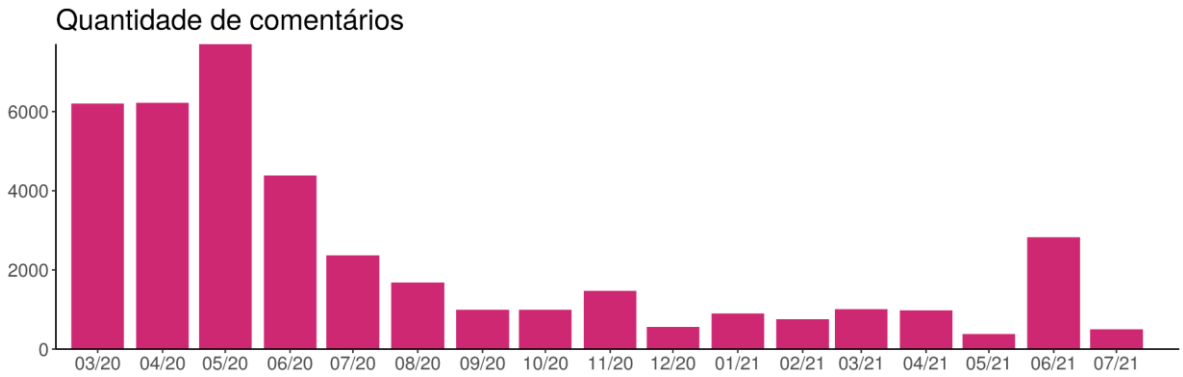

Figura 2 - Gráfico com a quantidade de comentários analisados por mês

Podemos observar na Figura 2 que a quantidade de comentários é maior nos primeiros meses da pandemia, resultado que verificamos como proporcional à maior quantidade de posts relacionados à pandemia nos primeiros meses da crise.

Os comentários passaram por um processo de preparação e limpeza, nas etapas:

1. Remoção de comentários duplicados
2. Remoção de URLs (endereços de páginas na internet)
3. Remoção de números

4. Remoção de menções (termos iniciados com @)
5. Remoção de acentos
6. Remoção de hashtags
7. Remoção de stopwords (palavras comuns da língua portuguesa)

Os comentários foram então tokenizados (divididos até a unidade de um único termo) e analisados com a técnica LDA com o algoritmo Gibbs Sampling.

No processo de modelagem, aplicamos a técnica para diferentes valores k que é a quantidade de tópicos desejada. Com o valor de k = 8, ou seja, 8 tópicos, obtivemos um conjunto de termos mais associados a cada tópico, e de comentários mais associados a cada tópico coerente com o cenário que conhecemos e que nos permitiu tirar algumas conclusões sobre o impacto da pandemia nos empreendedores.

Resultados e discussão

O resultado da modelagem com 8 tópicos pode ser observado no gráfico:

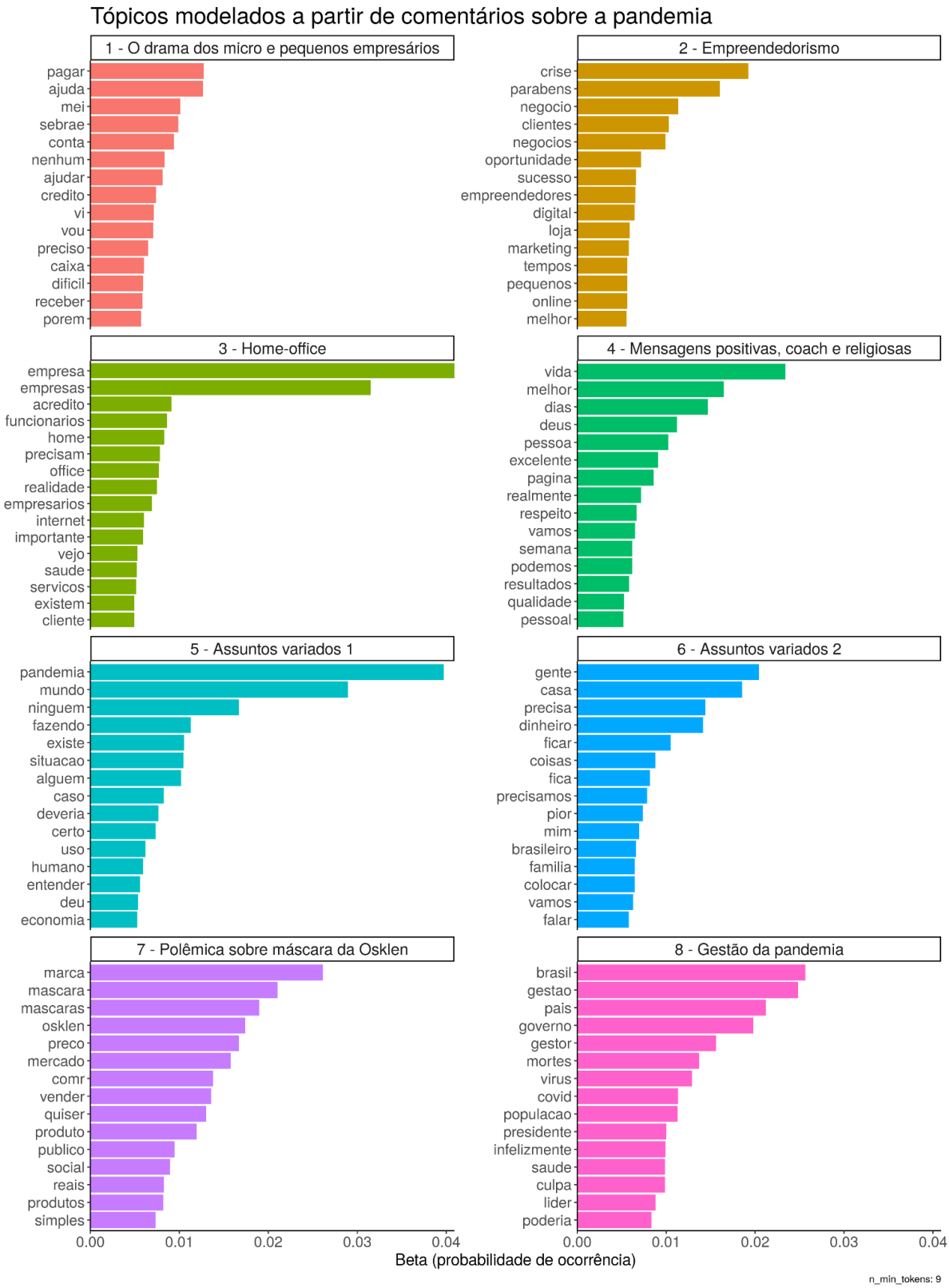

Figura 3 - Gráfico com resultado da modelagem exibe os termos mais comuns por tópico

Descrição dos tópicos modelados e identificados

Os tópicos modelados foram nomeados a partir da leitura dos termos de maior probabilidade (maiores betas), que podem ser observados no gráfico anterior, mas também foram nomeados principalmente a partir da leitura dos 20 comentários mais associados a cada tópico (maiores gammas).

Segue uma breve descrição dos tópicos a partir da análise dos comentários mais representativos de cada tópico:

Tópico 1 - O drama dos micro e pequenos empresários - março a julho/20

O impacto inicial da pandemia nos negócios. Comentários com reclamações sobre dificuldade de obter crédito, relatos dramáticos e pedidos de ajuda.

Tópico 2 - Empreendedorismo - abril/20 a fevereiro/21

Comentários com elogios ao Sebrae, relatos de experiências de superação, mensagens com orientações (coach empreendedor).

Tópico 3 - Home-office - abril/20 a junho/21

Comentários envolvendo a questão do home-office e a tensão desta mudança na relação entre empresas e colaboradores.

Tópico 4 - Mensagens positivas, coach e religiosas - fevereiro/20 a junho/21

Comentários com orientações (coach empreendedor), mensagens positivas de reflexão e de cunho religioso.

Tópico 5 - Assuntos variados 1 - março/20 a junho/21

Comentários sobre assuntos variados.

Tópico 6 - Assuntos variados 2 - março/20 a março/21

Comentários sobre assuntos variados.

Tópico 7 - Polêmica sobre máscara da Osklen - maio/20

Comentários em post do perfil “Administradores” chamando para artigo assinado que defende a posição da Osklen de venda de pack com 2 máscaras por R$ 147. Houve muitas críticas à postura da empresa, que se retratou, mas também houve defesas da empresa.

Figura 4 - Reprodução de post sobre máscara da Osklen. Publicado no perfil @administradores. Acessível em: https://www.instagram.com/p/B_3nnAvnD5a/

Tópico 8 - Gestão da pandemia - junho/21

Comentários em post do perfil “Administradores”, que questionou seus seguidores sobre qual a avaliação faziam sobre gestão da pandemia e se um bom gestor poderia ter evitado meio milhão de mortes. O post viralizou entre os apoiadores do presidente Bolsonaro que dominaram os comentários em sua defesa.

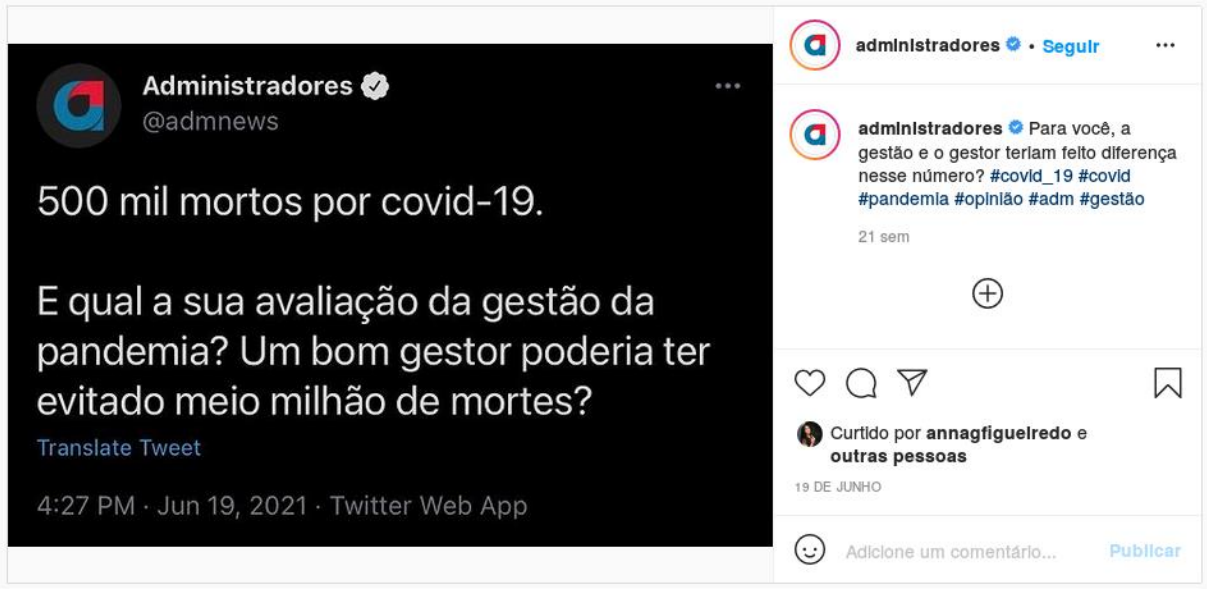

Figura 5 - Reprodução de post sobre gestão da pandemia. Publicado no perfil @administradores. Acessível em: https://www.instagram.com/p/B_3nnAvnD5a/

Assuntos identificados

Os 8 tópicos identificados nos permitem uma primeira leitura, porém alguns assuntos ficaram dispersos em mais de um tópico, como por exemplo os tópicos 5 e 6 que reúnem diferentes assuntos. Por isso também identificamos os principais assuntos a partir da leitura analítica dos 20 comentários mais associados a cada tópico:

Auxílio emergencial

O auxílio emergencial foi um dos assuntos mais recorrentes, em diferentes etapas, desde pedidos para sua criação até a liberação do benefício, indicando a importância do programa em um contexto de crise econômica. Comentários observados:

- Pedidos para a criação do auxílio
- Dificuldades no uso do sistema
- Pedidos para liberação do pagamento

Crédito

O crédito é mencionado principalmente pela dificuldade que os empreendedores encontraram para o acesso. Queixas observadas:

- Juros altos ou abusivos nos bancos privados
- Reclamação de burocracia para acesso a crédito a juros baixos nos bancos públicos (Banco do Povo, Pronampe, BNDES)

Sebrae

Os perfis do Sebrae foram bastante acionados com comentários sobre:

- Dúvidas variadas
- Pedidos para pressão no governo pelo Auxílio Emergencial
    - Pela criação do auxílio
    - Pela liberação do auxílio uma vez previsto
- Pedidos para pressão nos bancos pela prorrogação do prazo para o pagamento de dívidas
- Elogios
    - Pelas lives
    - Pelos cursos gratuitos oferecidos, como Marketing Digital

Home-office

O home-office trouxe uma nova realidade para as relações de trabalho e negócios que foi muito intensificada na pandemia. Listamos algumas observações e situações descritas nos comentários:

- Redução de custos com estrutura

- Valorização da educação à distância
- Se aplica bem a certas atividades
- Vantagem de acompanhar de perto o desenvolvimento dos filhos
- Dificuldades em conciliar o trabalho home-office e a maternidade
- Polêmica em torno da Lei 14.151/2021 que “Dispõe sobre o afastamento da empregada gestante das atividades de trabalho presencial durante a emergência de saúde pública de importância nacional decorrente do novo coronavírus”. Houve reclamações acerca de prejuízo aos contratantes;
- Melhor aproveitamento do tempo
- Melhor qualidade de vida
- Social presencial é importante mas não é necessário 100% do tempo presente no escritório
- Maior autonomia que requer boa gestão
    - Problemas com chefes sem maturidade, empatia ou confiança nas entregas
    - Equipes maduras entregam
- Nomadismo-digital, possibilidade de trabalhar de qualquer lugar
- Redução na qualidade de atendimento feito por profissionais em home-office

Cases de empreendedores

Nos comentários estão muitos relatos com histórias de dificuldade mas também de superação.:

- Profissional iniciou consultoria de inclusão digital de negócios (vendas online, delivery), aproveitando a demanda e aliando seus conhecimentos do ramo do comércio com conhecimentos do mundo digital

- Empresária se reinventou, a partir de uma live do Sebrae, fazendo cursos gratuitos, colocando em prática e obtendo sucesso com vendas online

- Empresário já tinha presença nas redes sociais e com a pandemia desenvolveu outra vertente profissional: consultoria para desenvolver o marketing digital para negócios do ramo alimentício

- A partir de 15 anos de experiência com eventos, em meio a pandemia, empresária cria agência especializada em realização de surpresas

- Vendedora de produtos de beleza buscou diversificação de atividade antes da pandemia com bolo gelado no pote. Durante a pandemia, com apoio do Sebrae, passou a fazer postagens que deram resultado

- Empreendedorismo social: Coordenadora de projeto social estava perdendo alunos porque foram despedidos ou tiveram salário reduzido. Decidiu fazer uma live com “dicas de como estudar inglês” em troca de contribuição a partir de R$ 10. Vaquinha quitou as dívidas dos alunos.

Caso Gabriela Pugliesi

No início da pandemia a influenciadora Gabriela Pugliesi publicou um story em seu instagram em que estava numa festa com amigos e dizia “foda-se a vida”. O incidente gerou muita polêmica nos comentários de post no perfil @administradores.

Economia vs. Saúde

Comentários acerca da dificuldade de honrar os compromissos com fornecedores e funcionários com o fechamento.

Drama de trabalhadores

- Desemprego e contas para pagar

- Dificuldade de conseguir emprego, auxílio emergencial ou alguma forma de conseguir dinheiro.
- Assumindo riscos de contrair o coronavírus para conseguir algum dinheiro

Conclusão

A modelagem de tópicos foi conclusiva permitindo a emergência de tópicos suficientemente definidos, o que facilitou a leitura para uma compreensão geral do impacto da pandemia. Dos 8 tópicos modelados, apenas 2 ficaram difusos, misturando diferentes temas. Para além do objetivo principal da pesquisa, alcançado principalmente na modelagem dos tópicos 1, 2 e 3, os demais tópicos permitiram uma compreensão mais ampla dos assuntos discutidos e comentados nos perfis no Instagram ligados ao universo do empreendedorismo e negócios.

A aplicação da técnica de modelagem de tópicos com LDA se mostrou eficaz e adequada para uma análise exploratória em um contexto de big data, permitindo uma amostra de comentários para análise entre dezenas de milhares, agrupada por tópicos, e sem os vieses de leitura da realidade se considerasse a análise apenas de conteúdos que nos chegam mediados pelos algoritmos das plataformas, como os conteúdos virais.

Referências

AGÊNCIA BRASIL. **Primeira morte por covid-19 no Brasil aconteceu em 12 de março**. Disponível em https://agenciabrasil.ebc.com.br/saude/noticia/2020-06/primeira-morte-por-covid-19-no-brasil-aconteceu-em-12-de-marco. Acesso em 14 março 2021.

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. **Latent dirichlet allocation**. the Journal of machine Learning research, v. 3, p. 993-1022, 2003.

ESHIMA, Shusei, KOSUKE Imai, TOMOYA Sasaki. **Keyword assisted topic models**. arXiv preprint arXiv:2004.05964, 2020.

FALEIROS, Thiago de Paulo et al. **Modelos probabilísticos de tópicos: desvendando o latent Dirichlet allocation**. 2016.

LU, Bin et al. **Multi-aspect sentiment analysis with topic models**. 2011 IEEE 11th international conference on data mining workshops. IEEE, 2011. p. 81-88.

PHAN, Xuan-Hieu; NGUYEN, Le-Minh; HORIGUCHI, Susumu. **Learning to classify short and sparse text & web with hidden topics from large-scale data collections**. Proceedings of the 17th international conference on World Wide Web. 2008. p. 91-100.

SILGE, Julia; ROBINSON, David. **Text mining with R: A tidy approach**. " O'Reilly Media, Inc.", 2017.

WICKHAM, Hadley; GROLEMUND, Garrett. **R for data science: import, tidy, transform, visualize, and model data**. " O'Reilly Media, Inc.", 2016.