RADAR AOS FATOS

Metodologia de redes • Out.2020

O que é o *Radar*?

O *Radar* **Aos Fatos** é uma ferramenta de monitoramento em tempo real do ecossistema de desinformação brasileiro. Nosso objetivo é reunir e classificar, de modo automatizado, conteúdos de baixa qualidade que circulam em sites e diferentes redes sociais, identificando com rapidez as publicações com potencial de viralização e a atuação dos grupos que operam na sua amplificação.

Compreender o fluxo da desinformação em múltiplas plataformas é fundamental em um cenário de crescente difusão de informação por meio das redes sociais. De acordo com o Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br), já em 2017, 73% dos brasileiros compartilhavam conteúdo na internet e 77% usavam redes sociais. Um relatório do International Center for Journalists (ICFJ) também aponta que os internautas no Brasil são mais propensos a receber notícias por redes sociais do que por outras mídias.

Para que um monitoramento de complexidade tão alta seja possível, desenvolvemos uma metodologia — em constante processo de aprimoramento — que alia conhecimentos da Linguística e da Comunicação à Ciência de Dados, implementada na linguagem de programação Python. Atualmente, o Radar monitora conteúdos publicados em **sites**, no **Twitter**, no **YouTube,** no **Facebook**, no **WhatsApp** e no **Instagram**.

A relevância do *Radar*

Dados do Reuters Digital News Report mostram que pela primeira vez desde 2013 as mídias sociais ultrapassaram a televisão em termos de consumo de mídia para notícias. **Facebook**, **WhatsApp** e **YouTube** são as redes mais usadas para esse fim. Isso não significa, no entanto, que os leitores confiem plenamente nas informações que recebem. O mesmo relatório mostra que 35% dos brasileiros se preocupam com a veracidade das notícias que recebem via **WhatsApp**, e 24% se preocupam com conteúdos do **Facebook**.

Estar apenas atento ao que circula nas redes não é suficiente para que empresas

que lidam com cenários de risco possam tomar boas decisões. Elas precisam estar continuamente conscientes de como as campanhas de desinformação influenciam no debate público, especialmente se esses debates estiverem sendo realizados em plataformas como **Twitter, WhatsApp e Facebook**.

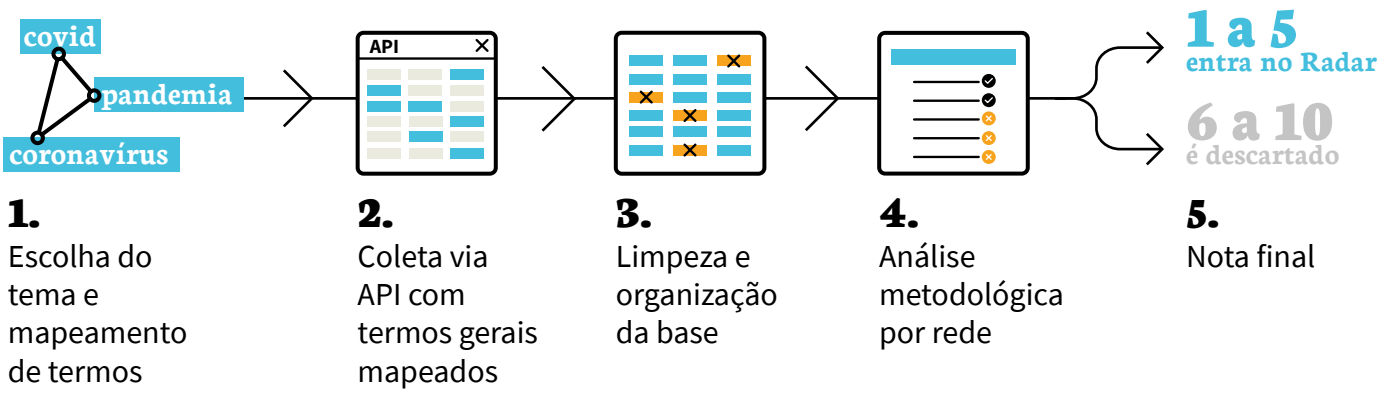
Existem muitas plataformas que acompanham o engajamento nas mídias sociais. A maioria delas, porém, concentra-se apenas na coleta de dados para análises de marketing. O que o *Radar* **Aos Fatos** oferece é a compreensão do amplo cenário de conteúdo de baixo nível, que tem impacto em grandes instituições e pode realmente afetar o cenário econômico e político.

Desde 2015 a equipe editorial do **Aos Fatos** monitora uma vasta gama de redes para descobrir o que é verdade e o que é falso e, mais importante, o que é relevante e o que não é. Sabemos o quão desafiador é usar várias ferramentas para acompanhar cada fluxo de conteúdo nas plataformas sociais. Para facilitar essa análise, criamos o *Radar* **Aos Fatos.**

Mas como funciona o monitoramento?

Em primeiro lugar, a equipe editorial do **Aos Fatos** seleciona um tema para monitoramento: inicialmente, por exemplo, o *Radar* teve como foco publicações sobre a pandemia de Covid-19. Em seguida, é preciso capturar as publicações em sites e redes sociais sobre o assunto escolhido. Após coletado, todo o material passa por uma série de processos que têm como objetivo extrair dados relevantes sobre conteúdo, autoria, imagens e vídeos, entre outras informações.

Nem todos os conteúdos coletados, contudo, serão exibidos no *Radar*. A seleção é feita por um algoritmo, com base em combinações complexas de termos de busca, que reúnem recortes do tema que representam maior risco de promover desinformação. No caso do coronavírus, por exemplo, publicações sobre medicamentos milagrosos ou sobre a origem do vírus têm maior potencial de disseminar desinformação do que conteúdos sobre o modo correto de usar máscaras ou relatos pessoais sobre o medo da pandemia. As combinações de termos são atualizadas diariamente.
Vejamos o seguinte tweet, a título de exemplo:



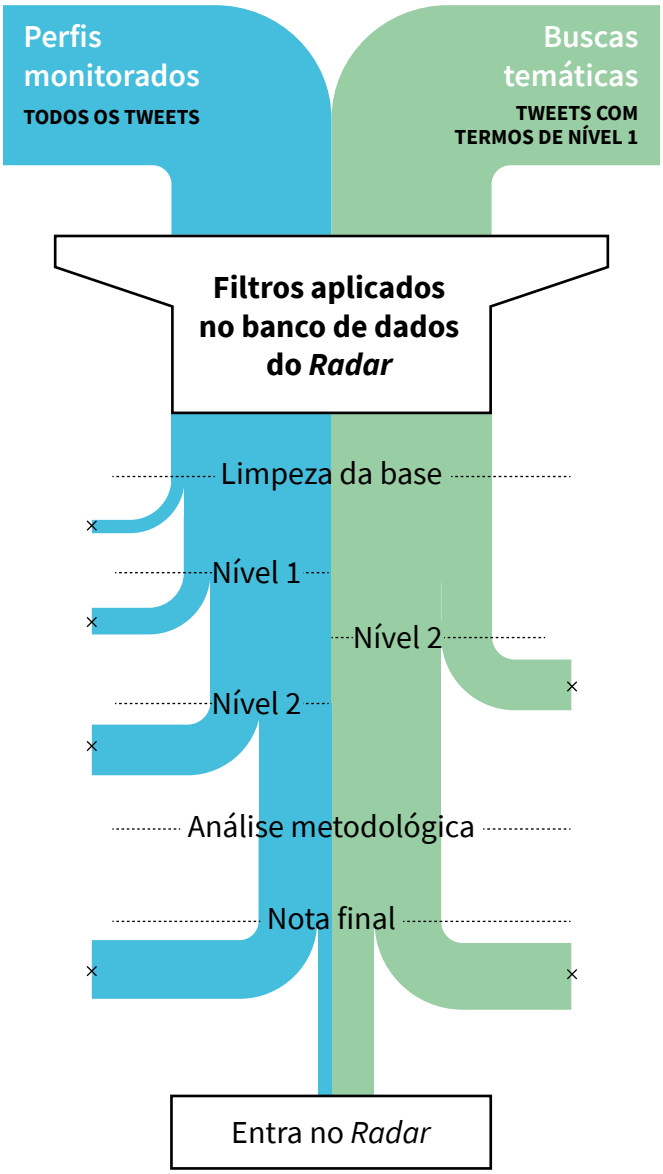
> ALERTA: Meu irmao, minha cunhada e os filhos dela estao curados de #Covid19. Todos tomaram kit covid. Nenhum deles precisaram ir pro hospital. Quais interesses estao por traz da demonizacao desse medicamento abencoado??? #CloroquinaSalvaVidas #CloroquinaSIM

No tweet em questão, podemos encontrar marcas do discurso que indicam um potencial da publicação estar compartilhando desinformação sobre o coronavírus. Tais padrões podem ser identificados não apenas por **hashtags**, como #CloroquinaSalvaVidas #CloroquinaSIM, mas também pelo uso de expressões e palavras da língua portuguesa que costumam ocorrer em textos que compartilham desinformação sobre coronavírus na web, como "estao curados", "kit covid", "quais interesses estao por traz", "demonizacao" e "medicamento abencoado". Tais termos, passíveis de mudanças e evoluções ao longo do tempo, são monitorados e atualizados constantemente pelo núcleo linguístico do *Radar*.

Além do uso de termos específicos à discussão sobre o coronavírus, também é possível identificar características que são comuns a conteúdos de baixa qualidade em geral: o recurso da **caixa alta** é um exemplo, bem como a presença de **erros gramaticais**. A palavra "alerta" também é outro sinal de que esta informação pode não ser confiável, já que é um indicativo de conteúdo com caráter alarmista. Outros indícios comuns de informação de baixa qualidade são, por exemplo, a presença de termos de natureza **ofensiva** ou **provocativa**.

Com os dados sistematizados, cada publicação é avaliada segundo um conjunto de

critérios, que recebem pontuações variáveis de acordo com um sistema de pesos. As notas compõem um índice que varia entre 1 e 10. Quanto maior o resultado, maior o nível de qualidade da publicação, ou seja, menores as chances do conteúdo analisado promover desinformação. **Apenas publicações com pontuação inferior a 5 são exibidas no *Radar***.



Cada plataforma analisada recebe um índice diferente, pois, apesar de algumas características em comum, há especificidades em cada uma delas que não podem ser ignoradas. Por exemplo, enquanto em **sites noticiosos** o uso de verbos ou pronomes de primeira pessoa é avaliado negativamente (pois foge da estrutura impessoal dos textos jornalísticos), essa mesma característica não é eficiente como critério para avaliação de publicações do **Twitter**, muito mais expressivas e pessoais.

Por outro lado, padrões temporais fazem mais sentido como critérios de avaliação no **Twitter** do que em sites; se um usuário desta rede publica muitas vezes em um intervalo curto de tempo — sinalizando uma possível ação coordenada —, ou se um retweet retoma uma publicação antiga — o que pode apontar descontextualização da informação — a publicação recebe nota inferior.

Sites

Embora muitas vezes simulem a estrutura e o estilo de notícias publicadas por veículos tradicionais de imprensa, os conteúdos noticiosos de baixa qualidade apresentam algumas características recorrentes que, analisadas em conjunto, contribuem para a identificação de publicações potencialmente enganosas.

Atualmente, são 36 os critérios considerados pelo *Radar* **Aos Fatos** na avaliação/classificação de conteúdos de sites segundo sua qualidade. Essas métricas são calculadas a partir de informações como: data e hora da publicação; local de publicação; autor(es); dados sobre imagens e vídeos; tags; estrutura textual do título e do texto completo, entre outras.

Critérios referentes ao conteúdo são os que representam maior peso na análise de sites. Entre os exemplos de características verificadas aqui, estão o uso de marcas de primeira pessoa, a preferência por expressões alarmistas e provocativas, os constantes erros gramaticais e o uso de padrões pouco usuais de pontuação.

Além de quesitos relativos ao conteúdo, são avaliadas, ainda, características referentes ao tempo (momento de publicação), à autoria, às mídias (imagens e vídeos) compartilhadas e aos perfis do site e/ou dos autores em redes sociais.

Twitter

As publicações de **Twitter** analisadas pelo *Radar* **Aos Fatos** têm origem em duas fontes de dados: o monitoramento de listas de perfis e as buscas temáticas, que utilizam combinações complexas de termos. Em ambos os casos, são analisados dados referentes tanto ao tuíte quanto ao usuário que o publicou. Ao todo, cerca de 40 critérios são considerados na análise de publicações do **Twitter**.

Para o monitoramento de perfis, duas listas foram organizadas: a primeira reúne usuários já identificados em ações coordenadas; que tenham características que se assemelham às de robôs; que já tenham sido identificados pela *Fátima* (@fatimabot) como disseminadores de desinformação ou que tenham mais de três tweets com pontuação inferior a 5 no *Radar*.

A segunda lista é composta por políticos com mandato em exercício, além de contas oficiais de governo e perfis de ministros e ex-ministros. Tais usuários são listados separadamente para que seja possível identificá-los em futuras análises, caso se deseje observar apenas atores do universo político em dado contexto.

Todos os tweets dos perfis presentes nas duas listas são gravados em um banco de dados. No entanto, serão exibidos no *Radar* apenas os tweets que se enquadrem nos temas

monitorados — atualmente, a Covid-19 e as eleições municipais de 2020. A seleção dessas publicações deve ser feita com um filtro de ao menos dois níveis.

Primeiramente, as publicações devem ser filtradas por um conjunto de palavras e termos de **nível 1** — mais amplas e, portanto, com resultados mais abrangentes. Os tweets filtrados neste primeiro nível passarão, ainda, por um segundo conjunto, de **nível 2,** onde estão incluídos diversos conjuntos de termos que se relacionam a assuntos recorrentes em publicações de baixa qualidade.

Não serão incluídos no *Radar*:

- Tweets com menos de 60 caracteres e sem mídia;

- Tweets que compilem ao menos quatro hashtags que estejam na lista de trending topics, uma vez que, de modo geral, representam uma compilação aleatória de temas diversos.

YouTube

A coleta de vídeos publicados no **YouTube** também parte de duas origens: uma lista de canais monitorados e as buscas temáticas. No primeiro caso, buscam-se conteúdos publicados por **canais reincidentes**, ou seja, que já compartilharam narrativas de desinformação ou que tenham obtido pontuação abaixo de 5 no *Radar* ao menos três vezes.

Já nas buscas temáticas, coletamos vídeos que contenham, no título ou na descrição, termos de pesquisa associados ao tema de interesse — no momento, a Covid-19 e as eleições municipais de 2020. Em seguida, filtramos tais publicações por um conjunto complexo de regras linguísticas que restringem o universo de dados a conteúdos com maior risco de serem de baixa qualidade.

Nos dois casos, cada publicação é avaliada em cerca de 30 critérios, que consideram pontos como o apelo a termos alarmistas e chamadas para ação, relações entre likes e dislikes e alterações específicas na grafia de palavras, entre outros. Além do vídeo em si, há métricas que observam características do canal publicador e mesmo dos comentários feitos por outros usuários sobre aquele conteúdo. Dados de engajamento são atualizados a cada 3 horas.

WhatsApp

A coleta de mensagens de texto, mídias e links no **WhatsApp** parte de uma parceria entre o *Radar* **Aos Fatos** e a Twist System, que nos auxilia no monitoramento de mais de 100 grupos públicos de discussão política. São grupos públicos aqueles cujos links de acesso são disponibilizados na rede por seus administradores. Atualmente, os grupos monitorados somam aproximadamente vinte mil mensagens por semana, que são filtradas de maneira automatizada e analisadas pelo núcleo linguístico do *Radar* **Aos Fatos**.

Na fase atual do *Radar*, são processados apenas os dados de texto e imagens, com extração de textos por meio de OCR (tecnologia de reconhecimento ótico de caracteres). Vídeos e áudios ainda não são analisados.

A categorização dos conteúdos atualmente coletados é feita a partir de filtros de palavras e expressões de primeiro e segundo níveis. O conjunto de nível 1 é composto por termos mais amplos relacionados ao tema principal como, por exemplo, “covid-19” e “coronavírus” ou “STF” e “judiciário”. Em seguida, com o conjunto de nível 2, filtramos as mensagens de modo que possamos, de maneira mais assertiva, identificar potenciais construções com desinformação ou qualidade baixa. Para isso, usamos 34 conjuntos complexos de palavras, que podem variar na medida em que novos termos ou assuntos surgem no Radar.

Cada mensagem coletada passa por 23 métricas de avaliação textual e 12 métricas de imagem, em sua maioria também utilizadas nas outras redes do *Radar* **Aos Fatos**. Especificamente para análises dos dados de **WhatsApp**, temos métricas que observam a recorrência no uso de emojis e a presença de estratégias discursivas que visam incentivar os leitores a se engajarem na propagação da mensagem recebida.

Futuramente, com a aplicação de métricas para contextualização que utilizam machine learning, serão acrescentadas dez novas métricas voltadas para análise de imagens.

Facebook

A coleta de publicações no **Facebook** é feita também a partir de dois caminhos: por uma lista de páginas monitoradas, compostas por páginas e grupos que já veicularam publicações checadas pelo **Aos Fatos**, e por buscas temáticas dentro de uma amostragem

de todo o universo de publicações da rede.

A coleta é realizada por meio do CrowdTangle, ferramenta do **Facebook** que é utilizada por empresas jornalísticas para monitorar conteúdos virais nesta rede social. O *Radar* analisa os posts recebidos a partir de mais de trinta critérios, incluindo parâmetros que analisam o conteúdo das publicações por uma combinação de fatores linguísticos. Textos inscritos em imagens compartilhadas também são avaliados dentro desses critérios, graças à tecnologia de reconhecimento ótico de caracteres (OCR) do CrowdTangle. Links compartilhados nessas publicações passam igualmente pela análise do *Radar*.

Os vídeos publicados no **Facebook** também são conteúdos coletados pelo *Radar*. Isso inclui mídias do **YouTube**, vídeos nativos da plataforma e o conteúdo completo de lives transmitidas pela rede social.

As regras implementadas incluem ainda a análise do potencial de viralização de cada publicação no momento da coleta e os tipos de reações manifestadas pelos usuários nos botões de engajamento da rede, como "*amei*" ou "*grr*". Os dados de engajamento são atualizados a cada 2 horas.

Instagram

O processo de coleta de publicações no **Instagram** é idêntico ao **Facebook**, já que as duas redes compartilham a mesma ferramenta, o CrowdTangle. Sendo assim, além das duas fontes de coleta, também temos acesso a textos inseridos em imagens, por meio da tecnologia OCR, disponível na plataforma.

O CrowdTangle permite coletar imagens, álbuns (publicações com mais de uma imagem), vídeos e IGTV (vídeos verticais). Atualmente, não são fornecidas imagens e vídeos no formato de "stories", que desaparecem após 24h.

Os dados extraídos ainda incluem a análise do potencial de viralização de cada publicação no momento da coleta, incluindo a relação entre o número de "curtidas" e comentários estimados pelo CrowdTangle e os que de fato a publicação obteve. Os dados de engajamento são atualizados a cada 2 horas.

Como é calculada a nota final?

O algoritmo que calcula as notas do índice de avaliação do *Radar* **Aos Fatos** foi elaborado com a aplicação de uma média ponderada das notas recebidas em cada critério, considerando os pesos de suas respectivas categorias. Por fim, foi realizada uma padronização das médias para deixá-las na escala de 1 a 10 de pontuação, por meio de uma interpolação, para cada publicação analisada.

A interpolação utilizou os valores de máximo e mínimo obtidos através da tabela de critérios de cada rede, isto é, são valores constantes. Isso permite que a adição de novas publicações, ao longo do tempo, seja totalmente adaptável à escala definida, e viabiliza a comparação entre conteúdos, independentemente do período de coleta.

NOTA	CHANCE DE CONTER DESINFORMAÇÃO	DESCRIÇÃO
1	Muito alto	Nosso algoritmo indica alta probabilidade do conteúdo apresentar erros, ser enganoso ou desinformativo
2	Alto	A publicação pontua mal na maioria dos critérios metodológicos: um conjunto de mais de 30 regras que avaliam conteúdo, autor e palavras-chave
3	Médio	Publicação pontua mal em regras críticas da metodologia, como por exemplo, a presença de termos alarmistas, de exagero ou generalização, entre outros, resultando em possível erro ou desinformação
4	Baixo	A publicação contém palavras-chave presentes em conteúdos desinformativos e não pontua bem nos critérios metodológicos relativos a conteúdo e autoria. Há chance do conteúdo ser enganoso
5	Muito baixo	Nosso algoritmo encontrou palavras-chave comuns a conteúdos desinformativos. Mesmo assim, o conteúdo pontua bem na metodologia e a chance de ser enganoso é baixa

Acurácia do monitor

Mensalmente o monitor de desinformação do *Radar* **Aos Fatos** é submetido a uma análise de precisão de acerto. O objetivo é avaliar se o algoritmo apresenta resultados satisfatórios e fazer eventuais melhorias para que ele seja mantido o mais exato possível.

Esse grau de precisão é dimensionado por meio de dois indicadores: a **faixa de acurácia**, que considera todas as publicações coletadas pelo monitor; e a **faixa de acerto**, que é calculada apenas entre as publicações que foram classificadas como de baixa qualidade.

Os dois cálculos são aplicados em todas as redes monitoradas (**Twitter, WhatsApp, Youtube, Instagram, Facebook e Web)**, e os resultados por cada tema monitorado são públicos.

Como é feita a análise

Uma vez ao mês a equipe do *Radar* gera uma amostra aleatória de publicações que foram processadas pelo monitor nos últimos sete dias e faz uma avaliação do conteúdo com olhar humano. O tamanho dessa amostra varia de acordo com o volume de publicações coletadas por tema e por rede (**Twitter, WhatsApp, Youtube, Instagram, Facebook e Web**). Cada amostra tem uma taxa de 95% de confiança e 5% de margem de erro.

Na análise humana, as notas que cada publicação recebeu de forma automatizada pelo monitor do *Radar* são desconsideradas, e a equipe avalia se cada um dos conteúdos deveria ser ou não classificado como de baixa qualidade. Em seguida, é feita uma comparação entre o resultado aplicado pela equipe e o que foi categorizado pelo algoritmo.

Por fim, são calculadas as porcentagens das publicações que entraram corretamente no *Radar* e daquelas que, por motivos metodológicos ou técnicos, não foram incluídas. Para a faixa de acurácia, são considerados todos os conteúdos que pontuaram de 1 a 10. Já a faixa de acerto é baseada apenas nas publicações que pontuaram de 1 a 5.

Como as publicações são avaliadas

Na comparação entre as avaliações humana e automatizada, a equipe do Radar classifica as publicações da amostra como **Falso Positivo (FP), Falso Negativo (FN) e Acerto (OK)**.

Falso Positivo (FP)

São conteúdos não desinformativos que receberam nota do algoritmo inferior a 5, ou seja, foram classificados como sendo de baixa qualidade. Isso pode acontecer por motivos como:

1. O conteúdo não é sobre o tema monitorado, mas acabou sendo filtrado pelas palavras-chave cadastradas, o que geralmente ocorre quando um termo tem diversos sentidos e se aplica em contextos diferentes. Nesses casos, o problema é resolvido com ajustes na regra de busca de forma que publicações com o termo em questão sejam coletadas apenas no contexto pesquisado.

2. Propagandas, correntes, piadas, sátiras e memes foram lidos pelo algoritmo de forma literal e classificados como desinformativos.

3. Uma publicação de alta qualidade apresenta algum marcador linguístico considerado característico de conteúdos desinformativos e, por isso, recebe uma nota mais baixa, em geral, próxima a 5. Durante a análise humana, o núcleo linguístico do *Radar* avalia se cada um desses casos deve ou não ser excluído do monitor.

4. Comentários, análises e textos de opinião de sites jornalísticos podem ser lidos como desinformativos quando tirados de seu contexto original. Um exemplo é quando textos de colunistas de veículos jornalísticos são replicados em outros sites ou em redes sociais sem os marcadores que inicialmente sinalizavam se tratar de uma opinião ou coluna. São desconsiderados como conteúdo de baixa qualidade textos analíticos ou de conscientização produzidos por jornalistas, pesquisadores, entidades científicas, acadêmicas ou de classe, como conselhos de medicina.

5. Conteúdos opinativos e de ataques políticos são classificados como de baixa qualidade quando as regras que balizam o algoritmo identificam palavras do léxico desinformativo nas publicações. Sendo assim, é considerado **Falso Positivo** o

conteúdo que, mesmo com esse tipo de léxico, não tem desinformação, marcas de exagero, alarmismo ou não usa termos ofensivos para atacar grupos ideologicamente distintos.

6. Mensagens de superação de doenças, condolências e luto devem ser analisadas pelas linguistas e, quando necessário, pela equipe de checagem de informações, uma vez que há grande circulação de conteúdo distorcido envolvendo curas e mortes relacionadas ao coronavírus. Caso não seja identificada desinformação no conteúdo, ele é excluído do monitor e classificado como **Falso Positivo**.

Falso Negativo (FN)
São conteúdos que possuem claramente marcações de desinformação dentro dos temas monitorados, mas que, por motivo técnico ou metodológico, não foram incluídos no *Radar* **Aos Fatos**. Ou seja, pontuaram acima de 5 ou foram excluídos. Em geral, eles são:

1. Publicações de baixa qualidade e que expõem, de forma articulada, ideias violentas, racistas, xenófobas, misóginas lgbtfóbicas e antidemocráticas que contenham desinformação. Nesses casos, é necessário atenção para diferenciar o conteúdo chulo (composto por xingamentos, ofensas pessoais etc) e opiniões controversas do que é essencialmente pertinente ao Radar.

2. Publicações que contenham algum nível de articulação textual com dados distorcidos e que provoquem a falsa sensação de informação confiável.

Acertos (OK)
São considerados acertos os casos de conteúdos cuja análise e cálculo do Radar correspondem à mesma análise e classificação feita por olho humano.

Como são feitos os cáculos

A classificação das publicações entre **Falso Positivo, Falso Negativo e Acerto** é tabulada em uma planilha na qual são discriminados dados como título, descrição de texto e/ou imagem e número de identificação da amostra de conteúdos coletados. Finalizada a revisão, os resultados são contabilizados e formatados como porcentagem. A partir destes resultados são, então, estabelecidas a **faixa de acurácia** e a **faixa de acerto.**

Faixa de acurácia
É o indicador que mostra a precisão do *Radar* **Aos Fatos** ao incluir ou excluir um conteúdo da análise. Para isso, ele considera o total das publicações da amostra (conteúdos que pontuaram de 1 a 10), não apenas as categorizadas como de baixa qualidade.

A faixa é determinada pelas porcentagens máxima e mínima dos conteúdos classificados como **"OK"** entre todas as redes e temas analisados. Assim, se a menor porcentagem for 62% e a maior 79%, teremos uma faixa de acurácia que varia entre esses dois pontos. Isso quer dizer que todas as publicações identificadas como equivalentes nas avaliações humana e automatizada figuram no intervalo entre essas duas taxas.

Faixa de acerto
É o indicador que mede o rigor da filtragem de conteúdo automatizada do *Radar* **Aos Fatos**, uma vez que avalia apenas as publicações que foram incluídas corretamente no monitor público por serem consideradas de baixa qualidade (pontuaram de 1 a 5).

A faixa de acerto é composta pela menor e pela maior porcentagem dos conteúdos classificados como **"OK"** entre todos os temas e redes monitorados. Se a menor taxa foi de 71% e o maior de 87%, a faixa de acerto será entre 71% e 87%, independentemente dos demais resultados.

Como os dados são expostos

Uma vez ao mês o *Radar* atualizará, em sua plataforma, a faixa de acerto calculada na amostra dos sete dias anteriores. Esse é o dado que demonstra o potencial de exatidão do *Radar* **Aos Fatos** quanto aos conteúdos monitorados em cada tema. A tabela com os resultados detalhados está disponível para o público e pode ser acessada clicando aqui.