UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

ROBERTO WALTER

# Caracterização e Comparação das Campanhas do Outubro Rosa e Novembro Azul no Twitter

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Prof[a]. Dr[a]. Karin Becker

Porto Alegre
2020

**CIP — CATALOGAÇÃO NA PUBLICAÇÃO**

Walter, Roberto

Caracterização e Comparação das Campanhas do Outubro Rosa e Novembro Azul no Twitter / Roberto Walter. – Porto Alegre: PPGC da UFRGS, 2020.

102 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2020. Orientador: Karin Becker.

1. Análise Demográfica. 2. Modelagem de Tópicos. 3. Avaliação de Similaridade. 4. Twitter. I. Becker, Karin. II. Título.

## AGRADECIMENTOS

É muito bom ter a quem agradecer ao final deste trabalho, que não reflete apenas anos de estudo e pesquisa, mas sim a conclusão de uma formação acadêmica. Primeiramente, agradeço a Deus que é Senhor, luz e direção em cada dia. Agradeço pela calma, tranquilidade e confiança de que tudo está correndo de acordo com Seu plano. Agradeço pelas oportunidades que encontrei e experiências que tive e que hoje formam o profissional que sou. Agradeço minha amada esposa Beatriz, a meus pais Arnildo e Astri, irmãs Mariely e Rubia, cunhado Joel, e o mais querido e pequeno novo membro da família Otávio, pelo apoio incondicional aos meus sonhos e anseios, além de orgulho que sempre demonstraram, e auxílio em cada pequena necessidade que tive. Agradeço a minha orientadora Dra. Karin Becker, por ser uma excelente orientadora e mentora que transmitiu muito de seu conhecimento teórico e prático. Ela que cuidou de detalhes, esperou por resultados e explicações, argumentos para fortalecer a teoria, mas também sempre se preocupou com meu bem estar, realização e satisfação no trabalho a ser feito. Obrigado por seu cuidado e encorajamento, além de seu incrível exemplo de pesquisadora. Agradeço também à Dra. Jonice Oliveira, Dra. Viviane Moreira e Dra. Renata Galante por participarem da banca de avaliação. Agradeço aos meus professores, colegas, amigos e companheiros no dia a dia. Agradeço ao Grupo DASS pela flexibilidade oferecida, em especial ao coordenador de TI Linos A. Kasper pela confiança e suporte durante minha jornada acadêmica. Ao Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande do Sul, pelo constante aprendizado e auxílio dispensados. Finalmente, agradeço às demais pessoas que me ajudaram direta e indiretamente na realização do trabalho e por falta de ajuda de minha memória, não relato aqui.

## RESUMO

As campanhas do Outubro Rosa e Novembro Azul buscam educar e conscientizar as pessoas sobre o câncer de mama e próstata. Apesar de sua semelhança, o Outubro Rosa é muito mais eficaz do que o seu homólogo Novembro Azul. Neste trabalho, contribuímos com uma análise comparativa do engajamento das pessoas nessas campanhas no Twitter. Primeiramente, usamos como base os tweets de 2017, caracterizamos a demografia dos tweeters, o nível de engajamento e o alcance das campanhas em 5 diferentes países. Encontramos diferenças e semelhanças nos padrões de atividade em campanhas e países. Em uma segunda etapa, estendemos a base de tweets entre os anos de 2014 a 2018, e realizamos uma avaliação dos períodos de atividade. Concluímos que ambas as campanhas do Twitter alcançam seu público-alvo. No entanto, as atividades do Novembro Azul estão concentradas no período de lançamento da campanha e seu alcance é mais limitado. Os tópicos trocados concentram-se na captação de recursos, enquanto o Outubro Rosa cobre mais assuntos.



**Palavras-chave:** Análise Demográfica. Modelagem de Tópicos. Avaliação de Similaridade. Twitter.

**Characterization and Comparison of the Pink October and Movember Campaigns on Twitter**


## ABSTRACT

Demographic Analysis, Topic Modeling, Similarity Assessment, Twitter

The Pink October and Blue November campaigns seek to educate and raise awareness about breast and prostate cancer. Despite its similarity, the Pink October is much more effective than its counterpart Blue November. In this paper, we contribute to a comparative analysis of people's engagement in these campaigns on Twitter. First, we used the 2017 tweets dataset, we characterize the demographics of tweeters, the level of engagement, and the reach of the campaigns in 5 different countries. We find differences and similarities in activity patterns in campaigns and countries. In a second step, we extend the base of tweets between the years 2014 to 2018, and perform an evaluation of the periods of activity. We conclude that both Twitter campaigns reach their target audience. However, Blue November's activities are focused on the launch period of the campaign and its reach is more limited. The topics exchanged focus on fundraising, while the Pink October covers more subjects.



**Keywords:** .

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| API | Application Programming Interface |
| LDA | Latent Direchlet Allocation |
| LSA | Latent Semantic Analysis |
| PLSA | Probabilistic Latent Semantic Analysis |
| CTM | Correlated Topic Model |
| NA | Novembro Azul |
| OR | Outubro Rosa |
| PLN | Processamento de Linguagem Natural |

**SUMÁRIO**

# 1 INTRODUÇÃO

Estudos estimam que uma em cada oito mulheres irá desenvolver o câncer de mama durante sua vida (ALTEKRUSE et al., 2010). A realização de exames de prevenção é fundamental para reverter este quadro. Hoje presente em vários países, a campanha anual do Outubro Rosa tem focado em aumentar a participação feminina nestes exames, bem como educar as pessoas e aumentar os cuidados em geral referentes ao câncer de mama (JACOBSEN; JACOBSEN, 2011; FOUNDATION, 2017). A campanha do Outubro Rosa teve início na década de 1990, quando o mês de outubro foi reconhecido oficialmente pelo governo dos EUA (Estados Unidos da América) como o mês de consciência sobre o câncer de mama. Esta campanha organiza caminhadas, eventos esportivos, distribui materiais de comunicação e incentiva o uso de vestimentas rosas. O principal objetivo é motivar mulheres a realizarem exames que detectam o câncer em estágios ainda iniciais, bem como angariar recursos financeiros. Estes esforços têm obtido muito êxito. Por exemplo, nos EUA o número de mulheres que realizaram exames preventivos subiu de 26% em 1987, para aproximadamente 72,4% em 2010 (CONTROL; PREVENTION, 2012). Um estudo (GLYNN et al., 2011) reporta que no mês de outubro há mais interesse no tópico de câncer de mama em pesquisas *online*. As campanhas do Outubro Rosa aumentaram o conhecimento da população permitindo um trabalho mais efetivo de prevenção, exames, conhecimento sobre tratamentos e pesquisas sobre o câncer de mama (EDGE, 2006). Isso indica que os trabalhos nas últimas décadas fruto destas campanhas de prevenção do câncer de mama têm surtido efeito.

O Novembro Azul é uma campanha com objetivos análogos aos do Outubro Rosa, com foco na população masculina para consciência e prevenção do câncer de próstata. Com a denominação de *Movember*, esta campanha surgiu na Austrália em 2003 em meio às comemorações do Dia Mundial de Combate ao Câncer de Próstata, celebrado em 17 de novembro. Ao longo do mês de novembro, o "mês da celebração do bigode", os homens são incentivados a deixar crescer o seu bigode com o objetivo de chamar a atenção para cuidados em geral da saúde dos homens, bem como angariar fundos para prevenção/cura do câncer de próstata.

Na prática, contudo, não se observa no Novembro Azul o mesmo nível de consciência e engajamento que ocorre no Outubro Rosa. Glynn et al. (2011) não observaram no mês de novembro um aumento das buscas usando termos relacionados ao câncer de próstata, a exemplo do que ocorre no Outubro Rosa em relação ao câncer de mama. Comparada com a feminina, a participação masculina em exames preventivos de câncer de próstata é ainda baixa (e.g. em uma campanha em 2011, apenas cerca de 20% do público masculino alvo (MCCARTNEY, 2012)). A inadequação das estratégias de conscientização utilizadas para o público masculino pode es-

tar contribuindo a estes baixos índices. Entender as diferenças do Novembro Azul em relação à sua contraparte bem-sucedida, o Outubro Rosa, permitiria identificar se existem diferenças relevantes nos padrões de envolvimento do público durantes essas campanhas. Esta compreensão permitiria desenvolver estratégias mais efetivas para abordar sua população-alvo, e aumentar a conscientização sobre a detecção precoce e tratamento do câncer de próstata.

A *internet* tem tido um papel vital na prevenção e tratamento de câncer. Aproximadamente 63% dos pacientes de câncer buscam informações oncológicas na *internet* (CASTLETON et al., 2011). Mais recentemente, trabalhos têm se dedicado a examinar a importância do uso de mídias sociais, em particular do Twitter[1], para propagar o conhecimento sobre a saúde pública (HIMELBOIM; HAN, 2014; LARANJO et al., 2014; BRAVO; HOFFMAN-GOETZ, 2016). O Twitter é uma mídia social de *microblogging* que permite aos usuários enviar e receber atualizações de outros contatos por meio de mensagens (*tweets*) de texto curtas, tipicamente limitadas a 140 caracteres. Para demonstrar seu engajamento em movimentos e causas específicas, usuários costumam utilizar *hashtags*. O Twitter tem sido ativamente utilizado para promover a companhas de Outubro Rosa (THACKERAY et al., 2013; NASTASI et al., 2017) e Novembro Azul (BRAVO; HOFFMAN-GOETZ, 2016; PRASETYO et al., 2015), através de *hashtags* tais como *#breastcancerawareness* e *#cancerdemama* (Outubro Rosa), e *#Movember* e *#NovembroAzul* (Novembro Azul).

Diferentes estudos contribuíram à análise das propriedades de campanhas sobre o câncer no Twitter. Em relação ao câncer de mama, Thackeray et al. (2013) avaliaram o volume dos *tweets* do Outubro Rosa de 2012 de acordo com diferentes assuntos, concluindo que as abordagens para comunicação e angariamento de recursos são apropriadas. Também identificaram que é necessário um melhor planejamento para que as celebridades e organizações tenham mais impacto no alcance das campanhas. Nastasi et al. (2017) avaliaram o volume de *tweets* sobre o câncer de mama em 2015, detalhado por demografia, localização e autoridade científica do perfil ou conteúdo, sugerindo esforços para a educação com informações precisas sobre o câncer. Já em relação ao câncer de próstata, Jacobson and Mascaro (2016) avaliaram o alcance dos *tweets* em uma campanha Movember em 2012, concluindo que o movimento Movember ainda sofre com um deficit de promoção do conhecimento e cuidados com a saúde. Prasetyo et al. (2015) chegaram a essa mesma conclusão ao avaliar o conteúdo dos *tweets* do Novembro Azul de 2013 para analisar o interesse em diferentes países em angariar recursos financeiros. Finalmente, Borgmann et al. (2016) comparam volume, localização, palavras e URLs mais frequentes sobre o câncer de próstata, testículo, bexiga e rim no Twitter em 2016, concluindo que o

[1] https://twitter.com

Twitter pode ser uma ferramenta promissora para influenciadores das campanhas sobre o câncer em geral.

Apesar destes trabalhos realizarem avaliações importantes para o entendimento do engajamento de usuários do Twitter nessas campanhas, ou o valor do Twitter para este propósito, existem ainda muitas lacunas. Em primeiro lugar, cada trabalho centra-se em algum tipo de câncer ou campanha específico (e.g. Outubro Rosa ou Novembro Azul), não permitindo a comparação direta entre o comportamento dos usuários em diferentes campanhas. Além disto, a análise dos comportamentos de engajamento está limitada em todos os trabalhos a um ano específico. Ainda, com raras exceções, não há uma análise estratificada por características demográficas dos usuários ou localização da campanha.

O objetivo deste trabalho é analisar o comportamento no Twitter relativo às campanhas do Novembro Azul (NA), comparado-o ao comportamento nas campanhas do Outubro Rosa (OR). Nossa pesquisa busca entender e comparar o engajamento dos usuários em termos de volume e propagação de *tweets*, as características demográficas dos usuários participantes, o comportamento ao longo de cada campanha, as características das campanhas em diferentes países, os principais assuntos abordados, bem como as variações deste comportamento ao longo dos anos. Em nossa análise buscamos responder as seguintes perguntas sobre o Novembro Azul, comparando-o com o Outubro Rosa:

- **QP 1**: Os usuários envolvidos nessas campanhas apresentam características de perfil demográfico e geográfico similares?
- **QP 2**: O comportamento ao longo dessas campanhas é similar?
- **QP 3**: As campanhas apresentam abrangência similar na rede social?
- **QP 4**: As campanhas abordam tópicos similares de conteúdo?
- **QP 5**: O comportamento das campanhas tem variado ao longo dos anos?

Para atingir estes objetivos, analisamos inicialmente 266.799 *tweets* coletados entre setembro e dezembro de 2017, definindo este comportamento como *baseline*. Posteriormente, comparamos as diferenças de comportamento do *baseline* de cada campanha em relação a anos anteriores (2014, 2015 e 2016) e o ano mais atual (2018). Para estes anos utilizamos o mesmo período de coleta do *baseline* (setembro a dezembro de cada ano), resultando em uma base completa de 1.387.497 *tweets*. Resultados preliminares desse trabalho foram relatados em (WALTER; BECKER, 2018).

Este estudo constitui uma experiência pioneira comparando campanhas de câncer no Twitter, com as seguintes contribuições:

- Complementamos estudos anteriores (THACKERAY et al., 2013; NASTASI et al., 2017; JACOBSON; MASCARO, 2016; PRASETYO et al., 2015; GLYNN et al., 2011; BORGMANN et al., 2016) com uma avaliação comparativa entre Outubro Rosa e Novembro Azul e com análises adicionais:

  ($i$) o público envolvido em cada campanha, detalhado em gênero e idade, verificando se corresponde ao respectivo público-alvo;

  ($ii$) avaliação de participação dos usuários em diferentes anos e em diferentes países, o que permite identificar comportamentos ao longo do tempo em diferentes contextos geográficos;

  ($iii$) os padrões temporais de atividades em cada campanha, detectando diferenças/semelhanças por campanha e país;

  ($iv$) diferenças na cobertura de tweets em cada campanha, mostrando que o Novembro Azul tem alcance limitado e que organizações e celebridades não desempenham um papel proeminente;

  ($v$) avaliação dos conteúdos discutidos durante as campanhas, mostrando que o Novembro Azul possui um foco diferente do Outubro Rosa. Essas análises podem ser estendidas a outras campanhas sobre câncer.

- Geramos métricas para uma campanha com resultados positivos (Outubro Rosa) e a contrastamos com uma campanha similar (Novembro Azul) com menos engajamento. Essa comparação ajuda a entender os fatores que influenciam o alcance da campanha do Novembro Azul no Twitter e, consequentemente, aumenta o engajamento da população-alvo em exames de detecção precoce do câncer.

- Estabelecemos um *baseline* com as campanhas no ano de 2017, e avaliamos sua evolução ao longo dos anos, identificando que houve mudança de foco nos últimos anos para as campanhas.

Este documento está organizado como segue. No Capítulo 2 é apresentado a fundamentação teórica das técnicas computacionais utilizadas. No Capítulo 3 são apresentados a motivação, trabalhos relacionados, similaridades e diferenças entre estes trabalhos e o aqui proposto. O Capítulo 4 define os materiais e métodos utilizados. Os experimentos realizados são

apresentados no Capítulo 5. Finalmente, o Capítulo 6 apresenta as conclusões e trabalhos futuros.

# 2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são apresentados de forma resumida os principais conceitos e técnicas utilizadas neste trabalho.

## 2.1 Modelagem de tópicos

Em mineração de textos é comum haver coleções de documentos como livros, sites de notícias, revistas ou *blogs*, os quais gostaríamos de separar em grupos. Por exemplo, identificar sites de notícias de esporte, livros sobre drama ou romance, *blogs* técnicos ou revistas de empreendedorismo. Modelagem de tópicos (CHANG et al., 2009) é uma área da mineração de textos para realizar a classificação não supervisionada de documentos em tópicos (assuntos). Um tópico representa um *cluster* de palavras que frequentemente ocorrem juntas em um conjunto de documentos.

Existem 4 métodos principais de modelagem de tópicos (ALGHAMDI; ALFALQI, 2015):

1. Latent Semantic Analysis (LSA) (DEERWESTER et al., 1990): o objetivo principal é criar conteúdos semânticos através de uma representação vetorial de textos;

2. Probabilistic Latent Semantic Analysis (PLSA) (HOFMANN, 1999): é uma evolução do LSA com objetivo de distinguir palavras relacionadas ao contexto, sem a utilização de um dicionário;

3. Latent Direchlet Allocation (LDA) (BLEI; NG; JORDAN, 2003): surgiu com o objetivo de melhorar os métodos PLSA e LSA em seus modelos de distribuição e probabilidades das palavras e documentos;

4. Correlated Topic Model (CTM) (LAFFERTY; BLEI, 2006): é uma evolução do LDA e utiliza distribuição normal logística para criar relações entre os tópicos.

O método CTM tem alta complexidade computacional pelos cálculos necessários, além de manter várias palavras comuns e genéricas dentro dos tópicos, o que pode dificultar sua avaliação (ALGHAMDI; ALFALQI, 2015). Assim, pelo fato de o LDA ser uma evolução dos métodos LSA e PLSA, e estar sendo muito utilizado por trabalhos similares ao nosso no propósito de classificar *tweets* (e.g. (THACKERAY et al., 2013; RESNIK et al., 2015; SURIAN

et al., 2016; ZHAO et al., 2011)), optamos por utilizá-lo neste trabalho para identificar os tópicos discutidos nas campanhas.

## 2.2 Latent Direchlet Allocation (LDA)

No campo de Processamento de Linguagem Natural (PLN) dentro da ciência da computação, Latent Direchlet Allocation (LDA) (BLEI; NG; JORDAN, 2003) é um modelo estatístico para ajustar e definir um modelo de tópicos a partir de um *corpus* linguístico. Ele trata cada documento do *corpus* como uma mistura de tópicos, onde cada tópico possui uma probabilidade de relação com o documento. Da mesma forma, cada tópico é composto por uma relação de palavras (ou termos) e probabilidade delas pertencerem ao tópico. LDA resulta em tópicos cujos termos têm maior probabilidade de co-ocorrerem juntos em documentos. Os tópicos possuem sobreposição de palavras, ao invés de serem conjuntos exclusivos e separados em grupos discretos.

O LDA recebe como entrada um *corpus* utilizado para o treinamento e descoberta de tópicos, e um parâmetro $k$ que determina a quantidade de tópicos a serem descobertos. Tipicamente, para melhoria dos resultados, este *corpus* passa por etapas de pré-processamento para limpeza do texto, como remoção de *stopwords*, remoção de caracteres especiais, *stemming*, etc (AGGARWAL; ZHAI, 2012; DENNY; SPIRLING, 2018). A saída do LDA é um conjunto de $k$ tópicos, os quais são compostos por palavras oriundas do *corpus* junto com a respectiva probabilidade $beta$ de pertencerem ao tópico. A saída também contém uma probabilidade $gamma$ de relação de cada tópico com cada documento do *corpus*. Assim, após descobertos os tópicos, o LDA descreve cada documento como uma distribuição dos tópicos encontrados. Por exemplo, dado três tópicos ($k = 3$), um documento pode ser descrito como sendo representado por 60% do tópico $k_1$, 20% do tópico $k_2$, e 20% do tópico $k_3$.

A Figura 2.1 apresenta um exemplo de aplicação do LDA. Neste exemplo, a aplicação foi realizada com a pré-definição por parte do autor para o LDA retornar 3 tópicos. Cada um dos tópicos é composto pela lista de palavras e a probabilidade de associação da palavra com o tópico. A imagem também contempla um gráfico de colunas com a probabilidade de associação de cada um dos tópicos com o documento. Como resultado, temos o tópico 1 com a maior probabilidade de relação com o documento. Algumas das palavras que compõem o tópico 1 são *car, vehicle* e *finance*, sugerindo que o documento trate do assunto de carros relacionado com uma situação financeira.

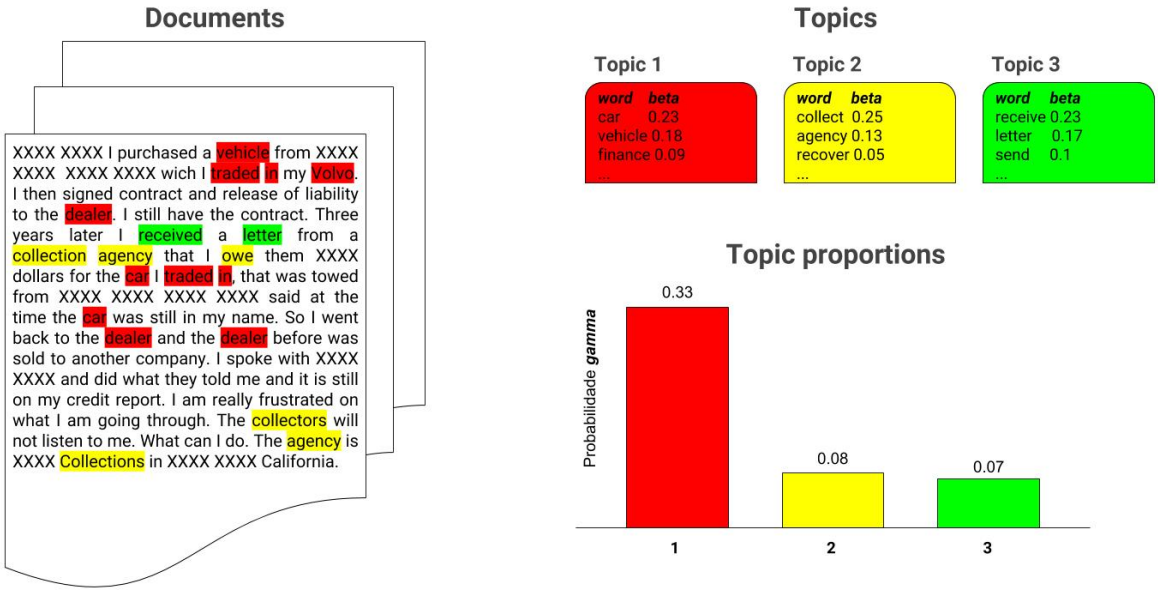


Figura 2.1: Intuições por trás do Latent Dirichlet Allocation.

Embora o conceito geral para utilização do LDA seja simples, o LDA apresenta alguns desafios na sua utilização, entre eles a definição do número de tópicos ($k$) a serem gerados, bem como a avaliação e interpretação dos resultados.

Medidas de avaliação dos resultados do LDA baseadas na distribuição de tópicos (e.g. W-Uniform, W-Vacuous e D-BGround (ALSUMAIT et al., 2009), e Método Bayesiano (MIMNO; BLEI, 2011)), produzem resultados úteis, mas são difíceis de serem interpretadas por humanos (RÖDER; BOTH; HINNEBURG, 2015). Métricas baseadas em coerência dos termos de cada tópico (RÖDER; BOTH; HINNEBURG, 2015), tais como CV (RÖDER; BOTH; HINNEBURG, 2015), purity (MANNING; RAGHAVAN; SCHÜTZE, 2010) e NMI (ESTÉVEZ et al., 2009), utilizadas para medir índices relacionados com a interpretabilidade de cada tópico, têm sido propostas para a avaliação da qualidade dos *clusters* resultantes do LDA.

Já quanto à interpretação do significado dos tópicos, é comum a inspeção manual dos termos pertencentes ao tópico de acordo com algum critério de representatividade, tal como sua probabilidade em relação ao tópico (MEHROTRA et al., 2013) ou valor de TF-IDF (frequência do termo–inverso da frequência nos documentos) (NETO; BECKER, 2018). Contudo este processo manual é subjetivo e inviável quando o número de tópicos é muito grande.

Neste trabalho, adotaremos métricas de coerência para determinar o número adequado dos tópicos ($k$). Também proporemos um método para interpretar um tópico em relação a termos de um referencial. Para isto utilizaremos representações de termos baseadas em contexto (*word embeddings*) e medidas de similaridade para viabilizar a interpretação de um grande número de tópicos.

## 2.3 Métricas de Coerência e Métrica $CV$

Existem diferentes métricas para avaliar a coerência da modelagem de tópicos (RÖDER; BOTH; HINNEBURG, 2015), como por exemplo, o Coeficiente de Correlação de Pearson usado em (RESNIK; GARRON; RESNIK, 2013), ou NPMI (Normalized Pointwise Mutual Information) adotada em (MEHROTRA et al., 2013). Röder, Both and Hinneburg (2015) propuseram uma métrica unificadora denominada $CV$, que combina diferentes dimensões para definição de coerência. Em uma extensa avaliação envolvendo sete métricas de coerência e diversos *benchmarks* (RÖDER; BOTH; HINNEBURG, 2015), CV foi a métrica que apresentou a melhor relação com a interpretabilidade dos resultados.

Uma medida de coerência é tratada como uma composição de partes que podem ser combinadas livremente, e agrupadas em dimensões que abrangem o espaço de configuração de medidas de coerência. Assim, a primeira dimensão é o tipo de segmentação usada para dividir um conjunto de palavras em subconjuntos. Estes subconjuntos são comparados uns contra os outros, por exemplo, segmentação em pares de palavras em que cada palavra é comparada com cada palavra do *corpus*. A segunda dimensão é o conjunto das medidas de confirmação que marca a concordância de um dado par de comparação, por exemplo, NPMI de duas palavras. A terceira dimensão é o conjunto de métodos utilizados para estimar as probabilidades das palavras para as medidas de confirmação, as quais podem ser calculadas de maneiras diferentes. Por exemplo, as medidas de confirmação *difference-* ($m_d$), *ratio-* ($m_r$) e *likelihood-measure* ($m_l$) (DOUVEN; MEIJS, 2007). Por último, a quarta dimensão é o conjunto de métodos de agregação em uma única métrica dos métodos de medidas de confirmação das probabilidades.

Em nosso trabalho, medimos o $CV$ do resultado do LDA utilizando diferentes tamanhos de $k$. Como mencionado, a métrica $CV$ retorna um índice de coerência dos tópicos gerados pelo LDA, e quanto maior este índice, maior a probabilidade de interpretabilidade dos tópicos da saída.

## 2.4 Word Embeddings

A forma de representação de palavras é um assunto importante para o processamento de linguagem natural (PLN). Representar as palavras como símbolos pode ser insuficiente para identificar possíveis relações semânticas e sintáticas entre as palavras ou outras atividades. Por exemplo, a representação simbólica das palavras "arroz" e "feijão" não são relacionadas, mesmo que "arroz" pode indicar uma ação do verbo "comer", não podemos inferir que "feijão" também

o seja. Por motivo dessas limitações, é buscado uma representação que captura similaridades semânticas e sintáticas entre palavras. Um desses paradigmas para obter este tipo de representação é *word embeddings* (LEVY; GOLDBERG, 2014). *Word embeddings* são uma representação do vocabulário de um documento no espaço vetorial. Cada uma das palavras é representada por um vetor de $D$ dimensões. Pela proximidade dos vetores no espaço, é possível identificar relações semânticas e sintáticas entre as palavras (PENNINGTON; SOCHER; MANNING, 2014), e assim, identificar o contexto de uma palavra no documento. Estes vetores podem ser manipulados no espaço em operações de média, distância, adição, subtração, etc.

Há diferentes formas de construção do espaço vetorial para *word embeddings*. Pode-se utilizar *embeddings* pré-treinados, ou gerar *embeddings* específicos a partir de um conjunto de documentos. A primeira abordagem é adequada quando não há um *corpus* de volume adequado para treinar o modelo, ou o conjunto de dados não pertence a um domínio específico. Já a segunda é pertinente quando se tem dados o suficiente para treinar o modelo e gerar os próprios *embeddings*, ou o domínio do problema é muito específico (QI et al., 2018). Alguns dos principais algoritmos para geração de embeddings são *word2vec* (GOLDBERG; LEVY, 2014) e *GloVe* (PENNINGTON; SOCHER; MANNING, 2014). Existem diferentes fatores que podem afetar a qualidade do *word embedding*. No trabalho de Schnabel et al. (2015) foram avaliados 6 métodos, onde o modelo CBOW do *word2vec*, teve desempenho melhor que outros métodos em 10 de 14 *datasets* utilizados no *benchmark*. O modelo CBOW (Current Bag-of-Words) (MIKOLOV et al., 2013), tenta predizer a palavra atual (central), baseado nas palavras ao seu redor, ou palavras de contexto.

### 2.4.1 Medidas de Similaridade

Medidas de similaridades e distância refletem o nível de proximidade entre um par de objetos. Uma variedade de medidas de similaridade e distância tem sido propostas e aplicadas, como distância Euclideana, coeficiente de Jaccard, semelhança de cosseno, coeficiente de correlação de Pearson e a divergência de Kullback-Leibler (RAJARAMAN; ULLMAN, 2011).

Quando documentos são representados no espaço vetorial, a similaridade de dois documentos corresponde com a correlação entre os dois vetores de representação. Essa similaridade entre dois vetores é quantificada como o cosseno entre os vetores, i.e., a semelhança de cosseno. A similaridade por cosseno é uma das mais populares medidas de similaridade aplicadas em identificação de semelhanças de documentos (BAEZA-YATES; RIBEIRO-NETO et al., 1999). Cada dimensão do espaço vetorial representa um termo com o seu respectivo peso no

documento. A métrica refere-se a uma orientação, onde dois vetores com a mesma orientação possuem uma similaridade de cosseno igual a 1, e dois vetores em orientação oposta (ou 180 de ângulo entre eles no espaço vetorial) possuem similaridade igual a 0. Essa métrica foi utilizada neste trabalho para medir a similaridade entre o vetor médio das *word embeddings* dos tópicos produzidos pelo LDA, e o vetor médio das *word embeddings* do nosso *baseline*, e a partir disso categorizamos de forma automática os tópicos produzidos.

# 3 TRABALHOS RELACIONADOS

Neste capítulo são discutidos os principais trabalhos de pesquisa relacionados à proposta deste trabalho.

## 3.1 Campanhas de propósito geral no Twitter

O uso das redes sociais para promover causas e campanhas tem crescido. O engajamento dos usuários nestas plataformas aumenta na medida que esses não são mais meros receptores passivos de conteúdo, mas sim membros ativos na geração e repasse de informação. O Twitter é uma das plataformas mais populares para postagem em tempo real de novidades, notícias, pensamentos e opiniões sobre os mais variados tópicos. Para muitos eventos, o Twitter tornou-se uma das principais fontes de notícias, atualizações e conscientização (LI et al., 2017).

Dados coletados a partir do Twitter têm sido utilizados para análise de diferentes causas e fenômenos sociais, como protestos políticos (LOTAN et al., 2011), violência de gênero (ELSHERIEF; BELDING; NGUYEN, 2017), equidade racial (CHOUDHURY et al., 2016; OLTEANU; WEBER; GATICA-PEREZ, 2016), impacto emocional provocado por fenômenos violentos em massa (HARB; BECKER, 2018), etc. Estes trabalhos buscam primariamente criar um *dataset* válido e representativo do contexto estudado. A maioria usa a API de integração do Twiter[1] para filtrar *tweets* que contenham palavras chaves ou *hashtags* em um período determinado de tempo. As análises temporais em geral são avaliações de volume de *tweets* e participação dos usuários durante períodos de campanhas *online*. O engajamento dos usuários tipicamente é medido pela atividade de postagem de *tweets* e *retweets*, detalhado no contexto de informações demográficas como sexo e idade, categorização de usuários e/ou geolocalização. Assim, busca-se definir as características de participantes e de perfis que possam ter mais ou menos influência/participação nos movimentos estudados.

Contudo, essas caracterizações demográficas e geográficas dos usuários apresentam desafios. Ao capturar o perfil de um usuário, o Twitter não solicita informações de sexo e idade, além de permitir que sua localização geográfica seja informada usando um campo aberto, que muitas vezes é preenchido com valores inválidos. Trabalhos têm adotado estratégias para extrair informações relevantes a partir dos dados informados no perfil dos usuários, tais como o Face++ (ELSHERIEF; BELDING; NGUYEN, 2017), Google Maps (MISLOVE et al., 2011), ou plataformas coletivas de definição de dados (OLTEANU; WEBER; GATICA-PEREZ, 2016).

[1] https://developer.twitter.com

O trabalho de ElSherief, Belding and Nguyen (2017) realizou uma análise em dados referentes às campanhas no Twitter de conscientização contra violência baseada no gênero, usando a *hashtag #notokay*. Para contornar o desafio de identificação das informações demográficas dos usuários, os autores propuseram utilizar a ferramenta Face++[2] que a partir da foto do perfil do usuário, estima a idade e o gênero. Experimentos realizados com esta ferramenta relatam uma acurácia de 85% (FAN et al., 2014). Esta estratégia também foi seguida por Harb and Becker (2018) para analisar emoções expressas no Twitter no contexto de ataques terroristas.

No tocante à determinação da localização geográfica, o Twitter permite que o usuário geo-referencie os seus *tweets*, mas o número de tweets geo-referenciados é muito baixo, tipicamente menor que 2% do volume total (SCHULZ; SCHMIDT; STRUFE, 2015). Assim, utiliza-se frequentemente a localização informada no perfil do usuário, ainda que esta informação possa ser imprecisa (e.g. o usuário está em local diferente do relatado no perfil, quando da postagem) e inválida por tratar-se de um campo aberto e sem validação. Esta estratégia foi inicialmente proposta por Sakaki, Okazaki and Matsuo (2010), sendo seguida por vários outros trabalhos (LOTAN et al., 2011; CHOUDHURY et al., 2016; HARB; BECKER, 2018). Neste caso, verificações de sanidade são empregadas utilizando listas de locais (HARB; BECKER, 2018) ou coordenadas no Google Maps (MISLOVE et al., 2011). Algumas plataformas de coleta de *tweets*, como Twitonomy[3] fornecem informações de localidade.

## 3.2 Análise de Campanhas Referentes ao Câncer de Mama e Outubro Rosa

Foram encontrados dois trabalhos de análise de campanhas sobre o câncer de mama no Twitter (THACKERAY et al., 2013; NASTASI et al., 2017), bem como um estudo que compara o interesse em pesquisas sobre câncer de mama e outros tipos de câncer (e.g. próstata) nos meses das respectivas campanhas (GLYNN et al., 2011). Estes trabalhos são discutidos no restante desta seção, com destaque para (THACKERAY et al., 2013), que embasa muitas decisões tomadas no presente trabalho.

O foco do trabalho de Thackeray et al. (2013) foi de analisar como o Twitter tem sido utilizado durante o mês de conscientização sobre o câncer de mama. Este trabalho avaliou o engajamento dos usuários de acordo com categorias pré-definidas de usuários, bem como em relação aos tópicos discutidos. Para o desenvolvimento do trabalho foram utilizados 1.744.271 *tweets* coletados entre 26/set/2012 e 12/nov/2012, por meio de consulta de palavras chaves na

[2] https://www.faceplusplus.com/
[3] www.twitonomy.com

Tabela 3.1: Tabela categorias e termos relacionados aos assuntos do Outubro Rosa definidos por Thackeray et al. (2013)

| **Categoria** | **Termos** |
|---|---|
| Wear pink | Wear, pink, socks, shirts, bracelet |
| Loved ones | Beat, survivor, grandma, mom, mamma, aunt, memory, die, story |
| Resentment | Other types, tired of, annoyed, resent, attention, fair |
| Walks & Runs | Walk, race, run, komen |
| Early detection | Mammogram(s), screening(s), exam, mammography, doctor, visit, detection, lump |
| Diagnosis | Symptom(s), diagnose(d), diagnosis |
| Treatments | Mastectomy, lumpectomy, chemo, adiation, chemotherapy, surgery, surgeon |
| Fundraising | Money, fundraiser, fundraising, donate, research, fund(s), donation, proceeds, benefit |

API do Twitter.

Em termos de categorias de usuários, Thackeray et al. (2013) classificaram os usuários do Twitter como entidades, celebridades ou usuários comuns, e quantificaram a participação dos mesmos no volume de *tweets*. Foram classificados como *organização* usuários cujo nome ou descrição de perfil contivesse palavras como *foundation*, *organization*, *agency*, etc[4]. *Celebridades* são perfis com conta verificada pelo Twitter com no mínimo 100.000 seguidores, e que não sejam classificados como organização. Todos os demais usuários foram categorizados como *Indivíduos*. Os autores concluem que a efetividade da propagação de informações por organizações depende de uma maior colaboração com os indivíduos e celebridades.

Thackeray et al. (2013) também avaliaram o conteúdo dos *tweets*, definindo o volume de *tweets* em tópicos sobre vestimentas, eventos de conscientização (*e.g.* caminhadas), exames para detecção precoce do câncer, testemunhos pessoais, diagnósticos, tratamento e angariação de fundos. Para a geração dos tópicos, foi utilizado o modelo probabilístico Latent Direchlet Allocation (LDA), sumarizado na Seção 2.2. Os autores não detalharam os parâmetros de entrada ou a forma de validação dos resultados. Segundo eles, os tópicos resultantes foram categorizados com um cruzamento manual entre as palavras associadas a cada tópico, e as palavras chaves definidas pelos autores constantes na Tabela 3.1. Por exemplo, se um tópico resultante do LDA contiver palavras como *mom, story* ou *survivor*, este tópico é categorizado como *Loved ones*, por essas palavras, neste contexto, geralmente estarem associadas a relatos sobre entes queridos.

Com o objetivo de identificar a confiança e veracidade dos *tweets* postados sobre a educação do câncer de mama, Nastasi et al. (2017) realizaram avaliações de volume de *tweets* sobre o câncer de mama entre novembro e dezembro de 2015. Para tal, avaliaram as características

[4]A lista completa é: cancer, foundation, foundation, health, department, organization, agency, news, group society, committee, volunteer, county, government, network, firm, company,companies, nurse blog, we, promotions, marketing, forum,campaign, pharmacy, and pharmaceutical

das principais URLs compartilhadas e as palavras mais utilizadas nos *tweets*, cruzando-as com informações demográficas, de localização e de autoridade científica dos perfis para propagar informações sobre o câncer de mama. Os autores alegam que as informações demográficas e de localização foram definidas com base nas características especificadas no perfil, mas não detalharam o procedimento utilizado.

Em termos comparativos, um estudo realizado por Glynn et al. (2011) analisou os níveis de atividades de busca no Google referente ao câncer de mama durante os anos de 2009 e 2014, comparando os resultados com os indicadores das buscas realizadas no mesmo período referentes ao câncer de próstata e câncer de pulmão. De modo geral, Glynn et al. (2011) identificaram que o câncer de mama incentiva mais buscas na *internet* comparado ao câncer de próstata e de pulmão. Os autores também observaram que durante o mês de outubro, mês de campanha de conscientização do câncer de mama, as buscas aumentam significativamente por termos relacionados ao câncer de mama. Este fato não foi observado para as buscas relacionadas ao câncer de próstata e os seus estímulos durante o mês de novembro para conscientização.

### 3.3 Análise de Campanhas Referentes ao Câncer de Próstata e Novembro Azul

No tocante ao câncer de próstata, identificamos trabalhos que realizam uma avaliação de engajamento e conteúdo propagado referentes a este tipo de câncer (PRASETYO et al., 2015; JACOBSON; MASCARO, 2016), bem como um trabalho que faz uma comparação de engajamento e conteúdo entre diferentes tipos de câncer, dentre eles o câncer de próstata (BORGMANN et al., 2016). Nesta seção realizamos a avaliação destes trabalhos.

Borgmann et al. (2016) examinaram diferenças em atividades, conteúdo, contribuintes e influenciadores por tipo de câncer (próstata, testículo, bexiga, e rim). A pesquisa abrangeu *tweets* relacionados às *hashtags* *#prostatecancer*, *#bladdercancer*, *#kidneycancer*, e *#testicularcancer* entre os anos de 2012 e 2014 no Twitter. A ferramenta Symplur[5], que mantém uma base de *tweets* relacionados a cuidados com a saúde, foi utilizada para a coleta avaliação de volume de *tweets*. O conteúdo foi avaliado através da ferramenta Tweet Archivist[6] pelo levantamento das 10 palavras mais frequentes e 10 *hashtags* mais relacionadas para cada *hashtag* avaliada. Os autores não especificaram o método de classificação dos usuários (e.g. *patient* e *doctor*[7]) e definição da geolocalização dos usuários para avaliar a distribuição global dos *tweets*, apenas

[5]https://www.symplur.com/

[6]https://www.tweetarchivist.com/

[7]A lista completa de classificações é: Patient, Doctor, Non-Doctor Healthcare Professional, Individual Other, Healthcare Organization, Organization Other, Spam

que essas informações foram fornecidas pela plataforma de análises Twitonomy[8], que retorna essas informações para os *tweets* dos últimos 6 a 9 dias. O trabalho identificou um crescimento de até 122% de 2013 para 2014 para as 4 campanhas no volume de *tweets*. A influência das classificações dos usuários foi medida em termos de número de menções e número de *tweets* postados, e o trabalho conclui que as organizações relacionadas à área da saúde são os principais influenciadores. Este foi um trabalho interessante de comparação das campanhas, que identificou que as discussões sobre as campanhas estão em uma crescente significante.

Prasetyo et al. (2015) analisaram o comportamento de campanhas do Novembro Azul no Twitter em diferentes países, com o objetivo de analisar como este comportamento está relacionado com a angariação de fundos à causa. Para isso, foi avaliada uma base de aproximadamente 2.000 *tweets* do ano de 2013, coletados pela API do Twitter. Os autores não utilizaram nenhum método computacional para categorização dos assuntos discutidos nesses *tweets*, desenvolvendo um trabalho de classificação totalmente manual. O trabalho avaliou e comparou estas atividades considerando campanhas no Canadá, Estados Unidos, Reino Unido e Austrália, países escolhidos por serem os 4 países de língua inglesa mais ativos nas campanhas de Movember no Twitter. Os autores concluíram que os usuários do Twitter focam principalmente nos aspectos sociais da campanha do Movember (e.g. crescimento do bigode). Há relativamente poucos *tweets* centrados em questões de saúde, e não há correlação significativas entre as atividades das campanhas e volume de doações.

Jacobson and Mascaro (2016) fizeram uma avaliação de atividades dos usuários ao longo da campanha do Novembro Azul de 2012, a qual foi detalhada por sexo. Os dados foram coletados utilizando a API do Twitter, no período de 31/out/2012 à 01/dez/2012, pelo filtro de algumas *hashtags* e usuários oficiais relacionados a campanha Movember. Os autores avaliaram o volume de participação dos sexos na campanha, mas o método de definição do sexo dos usuários não foi especificado. Os autores avaliaram as URLs mais compartilhadas pelos usuários, e constataram que as conversas são engajadas no sentido de divulgar a campanha como um movimento social, e não propriamente para tratar de questões sobre temas relacionados ao câncer de próstata, a exemplo do que acontece no Outubro Rosa (THACKERAY et al., 2013).

De modo geral, os trabalhos utilizaram períodos limitados de tempo para avaliação de atividades, o que impossibilita avaliar o comportamento das campanhas ao longo dos anos. Alguns trabalhos realizaram a avaliação de engajamento com a classificação do tipo de usuário, informações demográficas e de localização do usuário, mas este escopo nenhum dos trabalhos atendeu por completo. Ao contrário de Thackeray et al. (2013), discutido na Seção 3.2, nenhum

[8] www.twitonomy.com

destes trabalhos utilizou um método computacional para identificação de tópicos. Thackeray et al. (2013) avaliou apenas um ano da campanha Outubro Rosa, assim não sabemos se a posição da campanha é a mesma ou tem se modificado ao longo dos anos. Os trabalhos também não realizaram uma comparação de conteúdo na tentativa de identificar as direções que as campanhas mais recentes estão utilizando (e.g. Novembro Azul), em comparação com uma campanha mais antiga e com resultados consolidados (e.g. Outubro Rosa).

## 3.4 Considerações Finais

A Tabela 3.2 sumariza os principais aspectos destes trabalhos, e a diferença em relação ao nosso trabalho.

- Nosso trabalho estabelece uma análise comparativa da campanhas Novembro Azul e do Outubro Rosa, considerando diferentes anos e países. Os trabalhos comparativos entre diferentes tipos de câncer e envolvendo mais de um ano não focam em campanhas específicas e têm outros propósitos (GLYNN et al., 2011; BORGMANN et al., 2016);

- Assim como muitos trabalhos, consideramos o engajamento ao longo das campanhas, caracterizado pelo volume de postagens e padrões de propagação (*retweets*). Detalhamos esta análise considerando tipos, demografia, e localização dos usuários, detalhamento este que alguns trabalhos considerou apenas parcialmente, nem sempre especificando o método utilizado para identificação da demografia.

- Ao contrário dos trabalhos que consideram informações demográficas dos usuários (NASTASI et al., 2017; JACOBSON; MASCARO, 2016), extraímos estas informações automaticamente do perfil do usuário utilizando Face++. Adotamos as categoria de usuários propostas por Thackeray et al. (2013). Utilizamos o Google Maps como ferramenta para validação e retorno automático de uma localização válida a partir do perfil do usuário.

- Ao contrário dos trabalhos que analisam o conteúdo por termos/*hashtags*/URLs mais usadas (BORGMANN et al., 2016; NASTASI et al., 2017; JACOBSON; MASCARO, 2016; PRASETYO et al., 2015), extraímos os tópicos dos *tweets* automaticamente com técnicas de aprendizado não supervisionado (LDA). Diferentemente de Thackeray et al. (2013), que também usou LDA, desenvolvemos um método automático para avaliação dos *clusters* a fim viabilizar a interpretação dos tópicos no contexto de duas campanhas distintas e de diferentes anos.

- Os *tweets* coletados por nosso trabalho foram extraídos tanto através da API oficial do Twitter, para campanhas em andamento (2017, 2018), quanto utilizando outras formas de acesso para coleta de dados passados (2014-2016), não disponibilizados gratuitamente

Tabela 3.2: Visão geral de dados e métodos de trabalhos relacionados.

| **Trabalhos Relacionados** | **Contexto** | **Extração** | **Período** | **Engajamento** | **Tipo de Usuário** | **Nível Demográfico (Método)** | **Nível Geográfico (Método)** | **Análise de Conteúdo (Método)** | **Comparação entre campanhas** |
|---|---|---|---|---|---|---|---|---|---|
| Thackeray et al. 2013 | Outubro Rosa | API Twitter | 26/09/12 a 12/11/12 | Volume de tweets | Entidade Celebridade Indivíduo | | | Tópicos (LDA) | |
| Nastasi et al. 2017 | Outubro Rosa | Symplur | nov/15 a dez/15 | Volume de tweets | Autoridade científica | Sexo Idade (Não especificado) | Continente (Não especificado) | URLs Top-words | |
| Glynn et al. 2011 | Mama Próstata Pulmão | Google Insights for Search | 2009 a 2014 | Volume de Pesquisas no Google | | | | | Volume de buscas |
| Borgmann et al. 2016 | Próstata Rim Testículo Bexiga | Symplur, Tweet, Archivist, Twitonomy | 2012 a 2014 | Volume de tweets | Profissionais da saúde | | Continente (Twitonomy) | Top-10 palavras & Top-10 hashtags | Temporal Engajamento Tipo de Usuário Geográfico Conteúdo |
| Prasetyo et al. 2015 | Novembro Azul | API Twitter | 2013 | Volume de tweets | | | País (Auto-relatado) | Tópicos (Manual) | |
| Jacobson and Mascaro 2016 | Novembro Azul | API Twitter | nov/2012 | Volume de tweets | | Sexo (Não especificado) | | URLs | |
| Este Trabalho | Outubro Rosa & Novembro Azul | API Twitter & API Python | 24/set a 07/dez (2014, 2015, 2016, 2017, 2018) | Volume de tweets | Entidade Celebridade Indivíduo | Idade Sexo (Face++) | País (Auto-relatado & Google Maps) | Tópicos (LDA) | Temporal Engajamento Tipo de Usuário Demografia Geográfico Conteúdo |

pelo Twitter.

Em resumo, nosso trabalho é o único a realizar a comparação entre uma campanha consolidada e resultados comprovados (Outubro Rosa), e contrastá-la com uma campanha com propósito similar com público alvo diferente (Novembro Azul), mas que ainda está buscando a influência e importância do Outubro Rosa. Essa comparação foi realizada considerando diferentes países e anos, e detalhada de acordo com características dos usuários (categoria, idade e sexo), representando um estudo pioneiro no assunto. Ainda, ele atualiza os resultados de análises realizadas no início da década considerando dados mais recentes.

# 4 MATERIAIS E MÉTODOS

Neste capítulo são discutidos os materiais e métodos utilizados para as análises desenvolvidas neste trabalho.

## 4.1 Visão Geral

Os trabalhos relacionados apresentados no capítulo anterior utilizaram diferentes abordagens para análise de campanhas no Twitter relacionadas ao câncer. Nosso trabalho compara o comportamento dos usuários do Twitter ao longo das campanhas do Outubro Rosa e Novembro Azul sob a perspectiva das 5 questões de pesquisa (QP) definidas no Capítulo 1, as quais são relacionadas a atividades e perfil dos usuários (**QP1**), distribuição temporal dos *tweets* dentro do mês da campanha (**QP2**), abrangência e propagação das informações (**QP3**), análise de conteúdo (**QP4**), e seu comportamento ao longo dos anos (**QP5**).

Este estudo também contempla uma avaliação longitudinal deste comportamento. Primeiramente, realizamos as análises utilizando apenas o ano de 2017 a fim de detectar padrões de comportamento iniciais e estabelecer uma base de comparação. Posteriormente, comparamos estes padrões com anos anteriores (2014-2016), bem como com o ano seguinte (2018), visando observar semelhanças ou evolução. Portanto, coletamos dados tanto do presente, quanto do passado.

Parte das análises realizadas são baseadas em informações relativas ao perfil demográfico e geográfico dos usuários, juntamente com a categorização do mesmo (indivíduo, organização ou celebridade). Como estas informações não estão diretamente disponíveis, fizemos inferências a partir de dados disponíveis no perfil dos usuários do Twitter usando técnicas empregadas em trabalhos relacionados ElSherief, Belding and Nguyen (2017), Sakaki, Okazaki and Matsuo (2010), Thackeray et al. (2013).

Outra caracterização importante sobre as campanhas refere-se aos assuntos que os usuários compartilham no Twitter sobre as campanhas. Para isso, propomos um processo que envolve: a) a extração não supervisionada dos tópicos abordados usando LDA; b) a adaptação de um referencial existente (THACKERAY et al., 2013) para interpretação destes tópicos de acordo com categorias de assuntos; c) o mapeamento automático dos tópicos descobertos nas categorias pré-definidas, baseada na comparação de *embeddings* representando os *tweets* de cada tópico e as categorias disponíveis. Com isto, categorizamos de forma automática o assunto abordado em cada um dos *tweets*.

Em resumo, para desenvolver as análises de cada QP, necessitamos dos seguintes dados:

- **QP1**: caracterização do usuário de acordo com as seguintes dimensões:
  - Categoria, segundo (THACKERAY et al., 2013);
  - Sexo;
  - Idade;
  - País.
- **QP2**: distribuição temporal dos *tweets* ao longo do mês da respectiva campanha;
- **QP3**: identificação e separação entre *tweets* e *retweets*;
- **QP4**: tópicos (assuntos) representando o conteúdo dos *tweets*;
- **QP5**: caracterização do comportamentos detectados ao longo dos anos de 2014 a 2018.

O restante deste capítulo explica como os dados brutos foram obtidos, processados e categorizados, bem como os métodos utilizados para a investigação de cada questão de pesquisa. Como resultado produzimos uma base de dados com 1.387.497 *tweets* sobre as campanhas. O conjunto de dados utilizados neste trabalho está disponível em um repositório público[1], de acordo com as políticas de privacidade do Twitter vigentes.

## 4.2 Coleta dos Dados

As diferentes questões de pesquisa exigem a coleta de *tweets/retweets* nos respectivos períodos das campanhas (outubro/novembro), junto com as informações de seus usuários ao longo dos anos. O processo de coleta de dados iniciou-se pela campanha de 2017, quando coletamos os dados usando a API de consulta aos dados públicos do Twitter. A API retorna os *tweets* e *retweets*, com as seguintes informações: *text*, *favorited*, *favoriteCount*, *replyToSN*, *created*, *truncated*, *replyToSID*, *id*, *replyToUID*, *statusSource*, *screenName*, *retweetCount*, *isRetweet*, *retweeted*, *longitude*, *latitude*.

Como período de coleta, definimos o mês de cada campanha, precedido e seguido de uma semana. Isto resultou no período de 24 de setembro a 7 de novembro para o Outubro Rosa, e 25 de outubro a 07 de dezembro para o Novembro Azul.

[1] https://github.com/robertowtr/Campanhas-OutRosa-NovAzul-Twitter-Dataset

Definimos também um conjunto de termos de busca em diferentes línguas. Primeiramente, consultamos as *hashtags* indicadas pelo Twitter como relacionadas aos termos *Pink October* e *Movember*. Monitorando a coleta do ano de 2017, agregamos a este conjunto inicial algumas *hashtags* adicionais. O conjunto final de termos de busca está listado na Tabela 4.1, os quais contemplam estas campanhas nos idiomas inglês, português e espanhol, pois foram as de maior volume encontradas no Twitter.

Este processo foi então replicado para o ano de 2018, utilizando os mesmo termos de busca e o mesmo período. Como a API do Twitter permite apenas coletar dados da última semana gratuitamente, para os anos anteriores (2014-2016), utilizamos a ferramenta GetOldTweets-python[2] que permite a coleta de *tweets* sem restrição de período. No entanto, esse recurso possui a limitação de coletar somente *tweets*, e não *retweets*.

| Idioma | OR | NA |
|---|---|---|
| **Português** | #OutubroRosa, #PrevençãoContraoCâncerDeMama, #outubrorosabr, #cancerdemama | #CâncerDePróstata, #NovembroAzul |
| **Inglês** | #breastcancerawareness, #thinkpink, #pinkoctober, #walkagainstcancer, #breastcancer, #IDriveFor, @AmericanCancer, #breastcancerawarenessmonth, #projectpinkblue, #raceforthecure, #BCSM, #BRCA, #pinktober, #chokecancer | #Movember, #ProstateCancer, #BlueNovember, #beatcancer |
| **Espanhol** | #OctubreRosa, #miluchaesrosa, #luchacontraelcancerdemama | #noviembreazul |

Tabela 4.1: Conjunto de *hashtags*

No total foram coletados 1.387.497 *tweets* referentes a ambas campanhas, postados por 629.858 usuários diferentes. O volume total de *tweets* coletados para cada ano e campanha é mostrado na Figura 4.1.

Após a obtenção dos dados de *tweets*, que incluem o identificador do usuário que postou, utilizamos a API do Twitter para recuperar dados referentes aos usuários. Essa consulta retorna os seguintes atributos do usuário: *description, statusesCount, followersCount, favoritesCount, friendsCount, url, name, created, protected, verified, screenName, location, lang, id, listedCount, followRequestSent, profileImageUrl*.

[2]https://github.com/Jefferson-Henrique/GetOldTweets-python

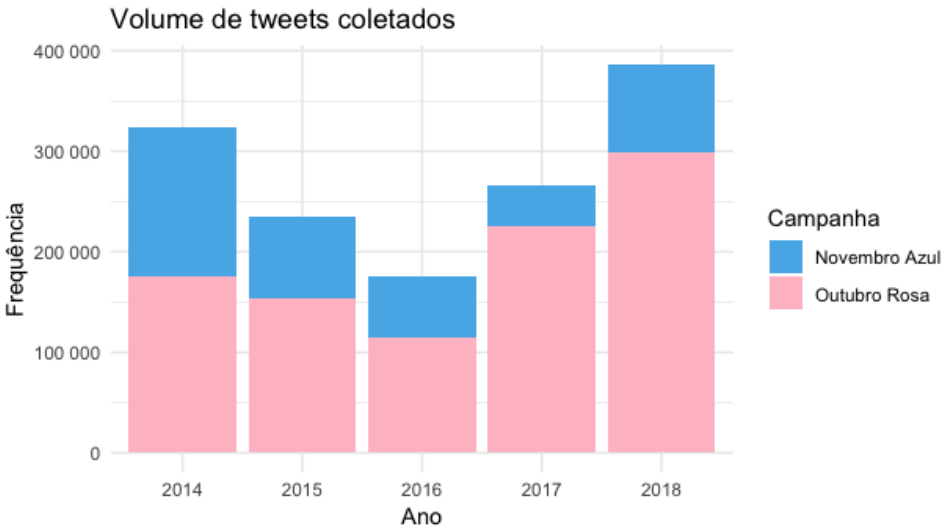


Figura 4.1: Volume total de *tweets* coletados para cada ano e campanha

### 4.3 Definição de demografia, categoria, e localidade dos usuários (QP1)

Para definirmos a demografia (sexo/idade), localização (país) e categoria de cada usuário, utilizamos dados contidos nos perfis dos usuários de todos os *(re)tweets* coletados.

Para definir a idade e gênero aplicamos o *Face++* sobre a foto coletada junto ao perfil dos usuários. Esta técnica foi proposta originalmente por Megvii (2013), e conforme reportado por Fan et al. (2014), experimentos com o Face++ apontam acurácia mínima de 85%.

Como mencionado, dado que o número de *tweets* geolocalizados é muito pequeno (menos de 2%), utilizamos o local definido no perfil do usuário, como originalmente proposto por Sakaki, Okazaki and Matsuo (2010). Para obtenção do país do usuário utilizamos a API do Google Maps[3] com a informação de localidade que o usuário preenche em um campo aberto no seu perfil.

Finalmente, classificamos os usuários de acordo com as categorias propostas por Thackeray et al. (2013), isto é, Celebridade, Organização, ou Indivíduo. Conforme explicado na Seção 3.2, para o perfil se enquadrar como Celebridade, ele deve ser verificado como verdadeiro pelo Twitter, possuir mais de 100 mil seguidores, e um gênero (masculino ou feminino) identificado pelo Face++. O usuário da categoria Organização é um perfil verificado pelo Twitter e que não possui o gênero identificado pelo Face++. Os demais perfis são classificados como Indivíduo.

[3]https://developers.google.com/maps/

## 4.4 Determinação dos tópicos dos tweets (QP4)

Para encontrar os tópicos abordados nos *tweets*, adotamos o modelo probabilístico Latent Direchlet Allocation (LDA) que foi apresentado na Seção 2.2. Contudo, como se tratam de diferentes campanhas ao longo de diferentes anos, existem dois grandes desafios neste processo. O primeiro é encontrar o valor mais apropriado de $k$, um dos parâmetros necessários ao algoritmo, para cada caso. Para este propósito, adotamos a métrica CV (Seção 2.3), que mensura a coerência dos tópicos encontrados.

O segundo desafio é a interpretação do significado de cada tópico. O resultado do LDA é um conjunto de documentos (tópicos), compostos por palavras e sua probabilidade de pertencerem ao tópico. Considerando o volume de tópicos a serem interpretados para o conjunto de anos/campanhas, a análise manual é inviável. Para automatizar este processo de análise, criamos um referencial de interpretação de tópicos que é uma adaptação dos tópicos descritos na Tabela 3.1 (THACKERAY et al., 2013). Então, usando uma representação de *word embeddings* e medidas de similaridade, comparamos as palavras mais relevantes de cada tópico com aquelas contidas no referencial de avaliação.

A Figura 4.2 apresenta o processo proposto para geração e avaliação dos tópicos, que é detalhado no restante desta seção. Devido à extensão das análises a serem realizadas e à dificuldade de interpretação dos resultados, limitamos a descoberta de tópicos aos *tweets* de língua inglesa.

### 4.4.1 Geração dos tópicos

Como esboçado na Figura 4.2, a entrada do processo é um *corpus* composto por todos os *tweets* de uma dada campanha para um dado ano (e.g. Outubro Rosa 2017). Sobre este são aplicadas as seguintes ações de pré-processamento:

1. Seleção dos *tweets* em inglês;
2. Remoção da referência de *retweet*;
3. Remoção de URLs;
4. Remoção de caracteres não alfabéticos;
5. Padronização do texto em minúsculo;

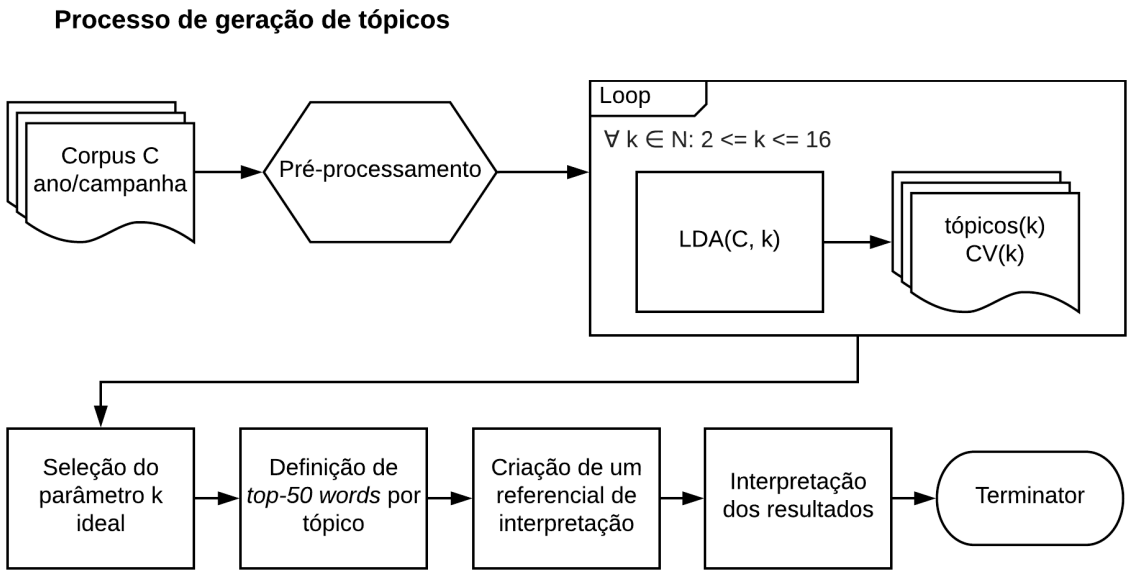


Figura 4.2: Fluxo do processo de geração de tópicos

6. Remoção de *stopwords*.

Na sequência, aplica-se o algoritmo LDA sobre este *corpus*. Para definição do parâmetro $k$ mais adequado ao *corpus* de cada campanha/ano, utilizamos a métrica $CV$, que avalia a coerência dos termos nos tópicos. Assim, executa-se o algoritmo utilizando diferentes valores de $k$, e para cada execução, calcula-se o coeficiente $CV$ correspondente. Tomadas todas estas execuções, seleciona-se dentre os três melhores valores de $CV$ aquele que mais se aproxima do número de tópicos do referencial de avaliação. Em outras palavras, o $k$ selecionado é uma aproximação entre o $k$ ideal indicado pela métrica $CV$ e o número de categorias utilizados para avaliar o seu significado. Como apresentado na próxima seção, criamos uma referencial a partir da adaptação do conjunto de categorias encontradas por Thackeray et al. (2013).

O Algoritmo 1 detalha as etapas para a definição do parâmetro $k$ de todos os pares campanha e anos. O parâmetro de entrada é o *corpus* completo, sendo este o conjunto de todos os *tweets* pré-processados. Para cada campanha/ano (linhas 4-5), são realizadas execuções do LDA variando o valor de $k$ entre 2 e 16 (linhas 7-8). Para cada uma destas execuções, é obtido o índice da métrica $CV$ para medir a qualidade do resultado do LDA (linha 9). Após essas execuções, selecionamos os 3 $k$s que apresentaram o melhor índice da métrica $CV$ (linha 11). Dentre estes 3 $k$s selecionados, optamos sempre pelo $k$ mais próximo da variável $nroCategoriasReferencial$ (linha 12), onde $nroCategoriasReferencial$ é definido pelo conjunto das 7 categorias que criamos no referencial da Seção 4.4.2. O retorno do algoritmo é conjunto de resultados do LDA para todas as campanhas e anos (linhas 13 e 15).

**Algorithm 1** Definição do parâmetro $k$

```
1:  procedure DEFINIR_PARAMETROS_K(corpus)
2:      result ← []
3:
4:      for [c, a] in (corpus.campanha, corpus.ano) do
5:          corpus_base ← filter(corpus, c, a)
6:
7:          for k = 2; k <= 16; k++ do
8:              result_lda[k] ← LDA(corpus_base, k)
9:              result_lda[k].cv ← CV(result_lda[k])
            end for
10:
11:         top3_k[] ← TopCV(result_lda, 3)
12:         k ← closest(top3_k, nroCategoriasReferencial)
13:         result[c][a].append(result_lda[k])
          end for
14:
15:     return(result)
```

| Campanha | Ano | Top 3 k's k (CV) | | | k selecionado |
|---|---|---|---|---|---|
| Novembro Azul | 2014 | 4 (0.443) | 2 (0.442) | 5 (0.403) | 5 |
| | 2015 | 2 (0.426) | 4 (0.414) | 8 (0.383) | 8 |
| | 2016 | 5 (0.421) | 3 (0.408) | 16 (0.400) | 5 |
| | 2017 | 11 (0.503) | 6 (0.460) | 2 (0.454) | 6 |
| | 2018 | 4 (0.399) | 12 (0.386) | 5 (0.373) | 5 |
| Outubro Rosa | 2014 | 3 (0.379) | 4 (0.372) | 5 (0.356) | 5 |
| | 2015 | 2 (0.396) | 16 (0.386) | 5 (0.381) | 5 |
| | 2016 | 3 (0.397) | 2 (0.388) | 11 (0.381) | 11 |
| | 2017 | 14 (0.391) | 15 (0.389) | 2 (0.348) | 14 |
| | 2018 | 3 (0.395) | 7 (0.387) | 10 (0.381) | 7 |

Tabela 4.2: Relação de campanhas e $k$-tópicos de entrada para o LDA

A Tabela 4.2 apresenta o resultado de seleção do $k$ para toda campanha/ano. Nela são indicados os 3 $k$s que tiveram o melhor índice na métrica $CV$. Dentre estes 3 candidatos, o $k$ mais próximo da variável $nroCategoriasReferencial$ foi selecionado como o parâmetro ideal de entrada do LDA para cada campanha/ano.

### 4.4.2 Criação de um Referencial de Interpretação

O resultado da execução do LDA para cada ano/campanha resulta em $k$-tópicos, cada qual representado por uma relação de palavras com a respectiva probabilidade de pertencer ao tópico. Para interpretar estes resultados, adaptamos os oito tópicos (categorias) encontrados no contexto de uma campanha de Outubro Rosa (Tabela 3.1). Nossa premissa é que as campanhas tratam do câncer prevalente em cada um dos sexos, e portanto esta transposição permite utilizar um referencial existente para câncer de mama no contexto do câncer de próstata. Assim, utilizamos as mesmas categorias para interpretar os tópicos das campanhas OR e NA.

Nossa adaptação do referencial descrito na Tabela 3.1 envolveu a inclusão de palavras que descrevem cada categoria, bem como a junção de duas categorias. Todas as adaptações propostas foram fruto da inspeção manual dos termos em amostras aleatórias de cada campanha, bem como em tentativas preliminares de interpretação com grupos de voluntários. Em resumo, fizemos as seguintes adaptações:

1. $ADP1$: Transpusemos algumas palavras para o contexto masculino do Novembro Azul (e.g. $mom$ vs. $dad$, $mammography$ vs. $PSA$).

2. $ADP2$: Após executar alguns experimentos, expandimos o conjunto inicial de palavras com termos que apareceram em *tweets* agrupados na mesma categoria de tópico, mas não estavam no conjunto inicial de palavras-chave daquela categoria. Por exemplo, identificamos as palavras $contribute$ e $charity$ em tópicos que incluíram a palavra $donate$, associada à categoria *Fundraising*, e estendemos essa categoria com duas palavras-chave adicionais. Outro exemplo são as palavras-chave $moustache$, $mustache$, $beard$ e $shave$ na categoria *Wear Blue*, uma vez que os homens geralmente deixam o seu bigode ou barba cresceram como forma de engajamento durante o mês da campanha. Estes símbolos são inclusive mais representativos que a cor azul como referência da campanha masculina.

3. $ADP3$: Unificamos as categorias *Early Detection* e *Diagnosis*. Em uma tentativa de análise manual destas duas categorias separadas, onde voluntários foram apresentados a nuvens de palavras característica de cada tópico, eles não conseguiram distinguir estas duas categorias.

4. $ADP4$: Adicionamos palavras que definem os 5 principais esportes nos EUA (futebol americano, basquetebol, hóquei no gelo, futebol e beisebol) e a sigla de seus respectivos campeonatos à categoria *Walks & Runs*, e modificamos o nome da categoria para *Walks &*

*Runs / Sports*. Adotamos essa medida sob a justificativa de que campeonatos esportivos apoiam a promoção destas campanhas.

O resultado dessas adaptações é descrito na Tabela 4.3, proposta para interpretação dos tópicos neste trabalho. As palavras em azul e rosa são resultado da adaptação $ADP1$, onde as palavras em rosa correspondem aos termos originais (Tabela 3.1), e as azuis, aos termos equivalentes no contexto do Novembro Azul. Palavras em rosa são utilizadas exclusivamente nos experimentos do Outubro Rosa, e da mesma forma, as palavras em azul são utilizadas exclusivamente nos experimentos referentes ao Novembro Azul. As palavras em negrito são aquelas acrescentadas como resultado da adaptação $ADP2$, e são utilizadas nas avaliações das duas campanhas.

Tabela 4.3: Categorias e palavras-chave para interpretação dos tópicos.

| **Categoria** | **Palavras-chave** |
|---|---|
| Wear pink/blue | wear, pink, blue, shirt, sock, bracelet,<br>moustache, mustache, shave, beard, noshave, shaving, stache |
| Loved Ones | grandma, mom, mamma, aunt,<br>grandpa, dad, papa, uncle,<br>beat, survivor, memory, die, story,<br>**friend**, **wife**, **husband**, **family**, **love**, **fight**, **hope**, **win**, **life** |
| Resentment | Other types, tired of, annoyed, resent, attention, fair |
| Walks & Runs / Sports | Walk, race, run, komen,<br>**football, soccer, basketball, hockey, icehockey, game, ball,**<br>**nfl, nba, nhl, mls, mlb** |
| Early Detection/Diagnosis | Mammogram(s), mammography, lump, PSA, rectal<br>screening(s), exam, doctor, visit, detection, prevent, early, check,<br>Symptom(s), diagnose(d), diagnosis |
| Treatments | Mastectomy, lumpectomy, prostatectomy, hormone,<br>chemo, radiation, chemotherapy, surgery, surgeon, treatment,<br>**amp**, **cure**, **patient** |
| Fundraising | Money, fundraiser, fundraising, research, fund(s), donate, donation,<br>proceeds, benefit, **contribute**, **help**, **charity**, **support(ing)**, **raise(s)** |

### 4.4.3 Interpretação dos tópicos

Observando os resultados da Tabela 4.2, é possível confirmar que a análise manual dos tópicos para cada ano/campanha é inviável, uma vez que há 29 tópicos referentes aos 5 anos do Novembro Azul, e 42 tópicos referentes ao mesmo período do Outubro Rosa. Portanto, propomos neste trabalho um método automatizado para interpretação de cada tópico baseado

na comparação das palavras mais relevantes do tópico e as palavras representativas de cada categoria em nosso referencial (Tabela 4.3).

Não consideramos a comparação simples de *strings*, pois dessa forma teríamos a limitação do método considerar somente as palavras previamente determinadas de forma manual em cada categoria do referencial, o que não permitiria valorizar sinônimos ou palavras similares às palavras determinadas nas categorias exceto se estas também fossem incluídas no referencial.

Com a utilização de *word embeddings*, por outro lado, podemos considerar estas palavras de acordo com seus contexto, e que geralmente aparecem junto a palavras pré-definidas nas categorias. Esse recurso permite que as relações e significados semânticos entre as palavras sejam ponderados. Além das relações semânticas e de contexto consideradas, *word embeddings* também providenciam uma representação vetorial das palavras, recurso este que utilizamos para identificar se um tópico gerado pelo LDA é similar ou não às categorias definidas no referencial. Mais especificamente, propomos comparar os vetores médios representando cada categoria e os termos com maior probabilidade de pertencer a cada tópico.

O Algoritmo 2 formaliza o conjunto de passos para esta categorização dos tópicos. Os parâmetros de entrada são o *corpus* completo pré-processado conforme descrito Seção 4.4.1, junto com os tópicos resultantes do Algoritmo 1 para cada ano/campanha. Primeiramente, treinamos o algoritmo *word2vec* com o $corpus$ de entrada e geramos *word embeddings* com 300 dimensões (linha 3). Em seguida (linhas 5 e 6), buscamos o *embedding* correspondente a cada termo do referencial de interpretação (Tabela 4.3), com as palavras gerais e as exclusivas de cada campanha (rosas para o Outubro Rosa, e azuis para o Novembro Azul). Então, com os valores dos *embeddings* das palavras individuais é calculado o vetor médio de cada categoria do referencial para o Outubro Rosa (linha 7) e o Novembro Azul (linha 8).

Em seguida, calculamos o vetor médio de cada tópico para cada campanha/ano usando as palavras mais representativas, usando os tópicos produzidos pelo LDA ($result\_lda$). Para todo tópico de cada campanha/ano (linhas 10-11), selecionamos as *top-50* palavras com maior probabilidade de pertencer a cada tópico da campanhas (linha 12). Para estas palavras, recuperamos os respectivos *embeddings* (linha 13) e os utilizamos para calcular o vetor médio do tópico (linha 14). Em seguida, calculamos a similaridade por cosseno entre o vetor médio do tópico e todos os vetores médios das categorias do referencial (linha 16). O resultado é o valor da métrica de similaridade e o índice da categoria do referencial mais similar ao tópico sendo avaliado. Caso a similaridade do tópico representado por este índice for igual ou superior a $0,5$, então o tópico é classificado com a categoria do referencial, caso contrário, ele é classificado como *Others* (linhas 18-22).

O resultado da aplicação deste algoritmo para todas as campanhas/ano são apresentados no Apêndice A.

### 4.4.4 Avaliação dos Resultados

A Figura 4.3 apresenta uma matriz com o resultado do Algoritmo 2 para a campanha do Outubro Rosa no ano de 2017, onde nas linhas são apresentadas as categorias do referencial, e nas colunas o identificador de cada tópico. Como exemplo, identificamos que o tópico 3 é mais similar à categoria *Early Detection/Diagnosis* ($similaridade = 0.807$), sendo classificado desta forma. Para conferir esta similaridade, geramos uma nuvem de palavras usando as *top-50* palavras com maior probabilidade de pertencer a este tópico, mostrada na Figura 4.4. Pode-se verificar que o tópico 3 está associado a palavras tais como *early, detection, screening, preven-*

**Algorithm 2** Definição de categoria dos tópicos

**procedure** CATEGORIZAR_TOPICOS(corpus, result_lda)

$word\_emb() \leftarrow word2vec(corpus, dim = 300)$

$top\_refer[] \leftarrow TopicosReferencial()$
$top\_refer[].words[].emb[] \leftarrow word\_emb(top\_refer)$
$vet\_med\_refer[OR][] \leftarrow VetorMedio(top\_refer, OR)$
$vet\_med\_refer[NA][] \leftarrow VetorMedio(top\_refer, NA)$

**for** `[c, a] in (result_lda.campanha, result_lda.ano)` **do**
  **for** `i = 2; i <= result_lda[c, a].k; i++` **do**
    $result\_lda[c, a][i].words[] \leftarrow TopWords(result\_lda[c, a][i], 50)$
    $result\_lda[c, a][i].words[].emb[] \leftarrow word\_emb(result\_lda[c, a][i])$
    $result\_lda[c, a][i].vetor\_medio \leftarrow VetorMedio(result\_lda[c, a][i])$

    $[idx, cosseno] \leftarrow MaxSimCos(result\_lda[c, a][i].vetor\_medio, vet\_med\_refer[c])$

    **if** $cosseno >= 0.5$
      $result\_lda[c, a][i].categoria \leftarrow top\_refer[idx].categoria$
    **else**
      $result\_lda[c, a][i].categoria \leftarrow$ "$Others$"
    **end if**

  **end for**

**end for**

$return(result\_lda)$

*tion, check, mammogram, mammography*, típicas da categoria *Early Detection/Diagnosis* como pode ser visto na Tabela 3.1.

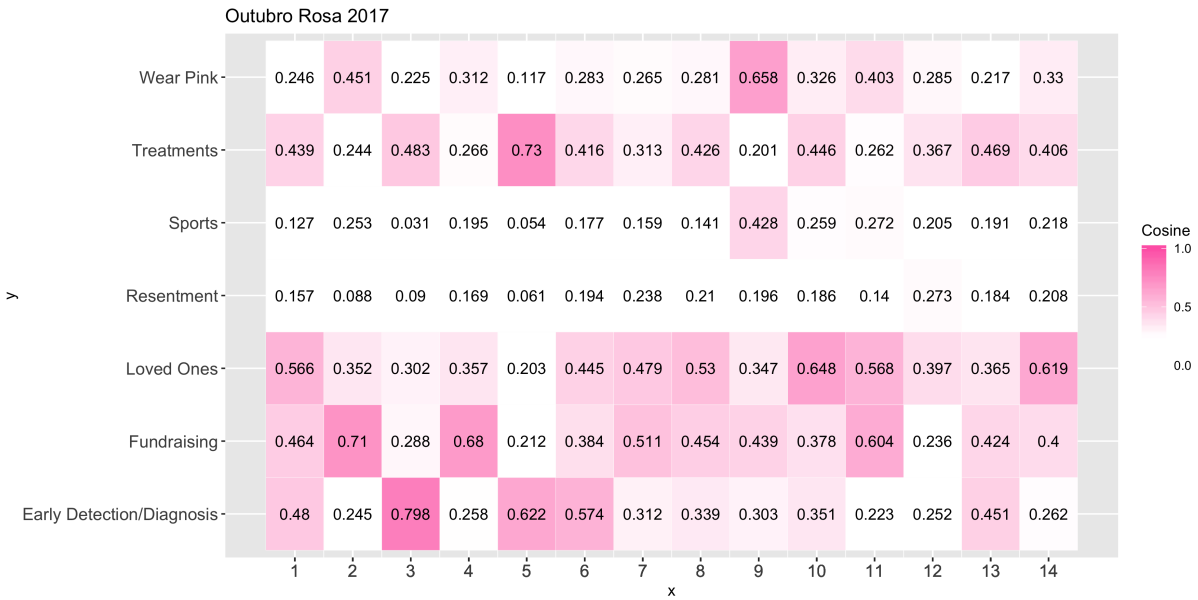


Figura 4.3: Matriz de similaridade da relação entre categorias do referencial e tópicos do Outubro Rosa do ano de 2017

Por outro lado, o tópico 12 da Figura 4.3 apresenta um índice baixo de similaridade com todas as categorias do referencial, sendo *Loved Ones* e *Treatments* as mais similares ($similaridade = 0.397$, e $0.367$, respectivamente), abaixo do nível de similaridade mínima definido ($i.e.$ 0.5). Ao avaliarmos as palavras deste tópico na nuvem de palavras da Figura 4.5, verificamos que não existe uma relação forte e clara com nenhuma das categorias do referencial na Tabela 4.3. Algumas palavras-chave como *contribute* e *amp* estão presentes na nuvem, mas apenas uma palavra de cada tópico não foi suficiente para atingir a similaridade mínima. Dessa forma, o tópico 12 é categorizado como *"Others"*.

Sobre a avaliação do Novembro Azul, a Figura 4.6 apresenta a relação de resultados para a campanha do ano de 2017. Por exemplo para o tópico 1, mais similar à categoria *Wear Blue* ($similaridade$ = 0.715), vemos a presença de palavras como *mustache, beard, stache*, e *shave* na nuvem de palavras correspondente, mostrada na Figura 4.7.

A relação completa de matrizes de similaridade para ambas as campanhas em todos os anos está disponível no Anexo A. Já o Anexo B apresenta as nuvens de palavras de todos os tópicos assim extraídos. O autor realizou um trabalho manual de inspeção da sanidade entre


Figura 4.4: Wordcloud do tópico 3 da campanha do Outubro Rosa de 2017


Figura 4.5: Wordcloud do tópico 12 da campanha do Outubro de 2017

a categoria mais similar e as palavras das diferentes nuvens, concluindo que os resultados são coerentes.

A Tabela 4.4 apresenta o resultado geral deste processo, com indicação de aderência dos tópicos às categorias do referencial ou não (neste caso, classificados como $Others$). A coluna $k$ indica o número de tópicos extraídos da campanha/ano de referência na linha. As colunas $Referencial$ e $Referencial(\%)$ indicam o número absoluto e de proporção de tópicos classificados em uma categoria do referencial. Da mesma forma, as colunas $Others$ e $Others(\%)$ indicam quantos tópicos foram classificados como $Others$, ou seja, tópicos que não tiveram similaridade com nenhuma categoria do referencial.

Quando verificamos os resultados da campanha do Outubro Rosa, identificamos que o método resultou em uma classificação com boa cobertura em relação ao referencial, em um

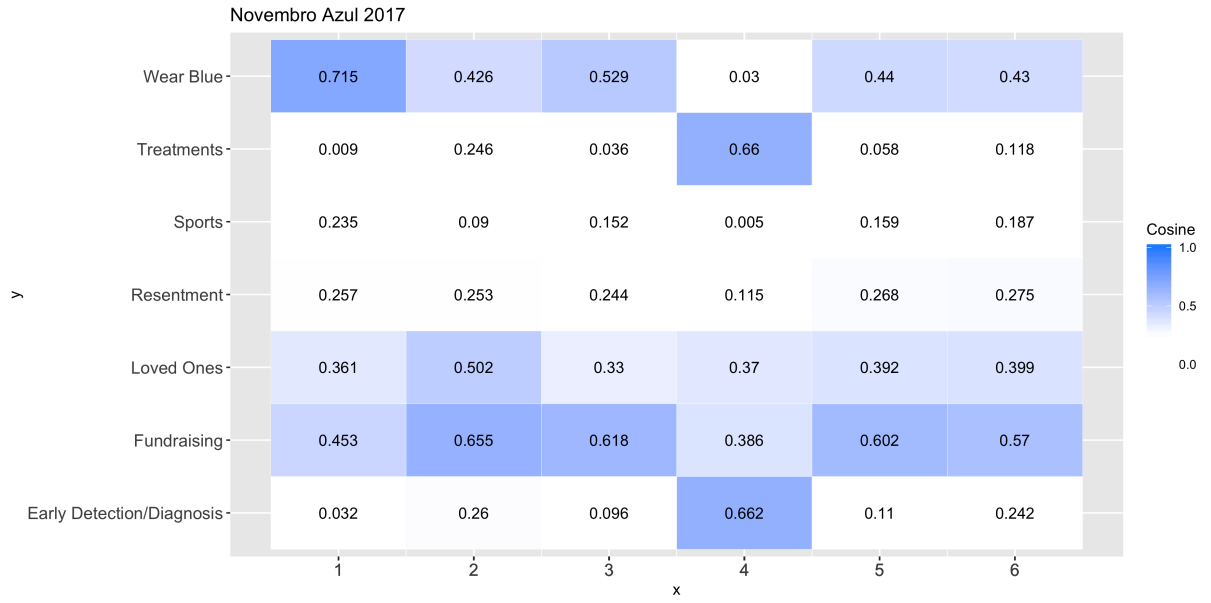


Figura 4.6: Matriz de similaridade da relação entre categorias e tópicos do Novembro Azul do ano de 2017

total geral de 81% de tópicos classificados. Para a maioria dos anos, o método apresentou um índice de categorização de no mínimo 80%. Apenas um ano ficou abaixo, mas ainda assim teve mais classificações de categorias do referencial (64%), comparado à categoria $Others$ (36%). Estes resultados nos indicam que o método proposto funcionou bem com o referencial definido. O Novembro Azul apresentou um índice de classificação do referencial ainda mais expressivo que o Outubro Rosa, com um total geral de 93%. Em especial, nos últimos 3 anos da campanha (2016, 2017 e 2018) o método classificou 100% dos tópicos. O menor índice ficou para o ano de 2014, com 80% dos tópicos categorizados.

Os resultados positivos obtidos tanto para o Outubro Rosa, quanto para o Novembro Azul, indicam a capacidade de estender o referencial e o método de categorização automática para outras campanhas, como por exemplo, o Setembro Amarelo que busca conscientizar as pessoas para não cometerem suicídio. Com a modificação de algumas palavras chaves peculiares a cada campanha, o referencial pode ser rapidamente adaptado e o método se encarrega, com base nas relações semânticas, de identificar palavras importantes em novas campanhas avaliadas. Além disso, a categorização automática de tópicos permite que as análises sejam facilmente escaladas.


Figura 4.7: Wordcloud do tópico 1 da campanha do Novembro Azul de 2017

Tabela 4.4: Resultado do método de categorização dos tópicos.

| Campanha | Ano | k | Classificação | | | |
|---|---|---|---|---|---|---|
| | | | Referencial | Referencial (%) | Others | Others (%) |
| Outubro Rosa | 2014 | 5 | 5 | 100% | 0 | 0% |
| | 2015 | 5 | 4 | 80% | 1 | 20% |
| | 2016 | 11 | 7 | 64% | 4 | 36% |
| | 2017 | 14 | 12 | 86% | 2 | 14% |
| | 2018 | 7 | 6 | 86% | 1 | 14% |
| | **Total** | **42** | **34** | **81%** | **8** | **19%** |
| Novembro Azul | 2014 | 5 | 4 | 80% | 1 | 20% |
| | 2015 | 8 | 7 | 88% | 1 | 13% |
| | 2016 | 5 | 5 | 100% | 0 | 0% |
| | 2017 | 6 | 6 | 100% | 0 | 0% |
| | 2018 | 5 | 5 | 100% | 0 | 0% |
| | **Total** | **29** | **27** | **93%** | **2** | **7%** |

# 5 EXPERIMENTOS

Neste capítulo são apresentados os experimentos para as questões de pesquisa (QP) definidas no Capítulo 1. Os experimentos das QP1-QP4 são baseados nas campanhas do ano de 2017. Os resultados de evolução das campanhas nos demais anos são discutidos na seção da QP5.

## 5.1 QP 1: Os usuários envolvidos nas campanhas apresentam características de perfil demográfico e geográfico similares?

Para responder esta pergunta, utilizamos os dados extraídos do perfil do usuário para determinar seu gênero, idade, país e classificação do usuário. Ao longo desta seção, detalhamos como estas características definem o comportamento de cada campanha ao longo do ano de 2017.

Na Figura 5.1 pode-se visualizar a distribuição dos *tweets* por gênero e campanha sobre o total de *tweets*. Identifica-se que cada gênero engaja mais na campanha da qual é alvo, i.e., mulheres no Outubro Rosa e homens no Novembro Azul. A Figura 5.1 também mostra que há uma participação muito pequena de organizações (1.22% no NA e 1.11% no OR), mostrando que existe ainda um espaço bastante grande de crescimento nesta categoria de usuários.

Avaliamos em seguida como esta participação ocorre em diferentes países. Para este fim, selecionamos os 5 países com maior volume de *tweets*, a saber, Brasil, Estados Unidos, Reino Unido, Canadá e México. A Figura 5.2 apresenta a distribuição de gênero e organizações para estes países. Pode-se verificar que com exceção do Canadá, em todos os países a campanha do OR atrai maior participação quando comparada a do NA. Os Estados Unidos e o Brasil são os países que apresentam as maiores diferenças, onde a participação no NA está abaixo dos 10% nos dois países.

As organizações representam uma taxa muito baixa de participação em todos os países. Pode-se observar uma pequena participação de organizações no OR no México (3.92%) e EUA (2.11%). Quanto ao NA há um envolvimento observável apenas no Canadá (2.52%) e México (1.42%). Nos demais países a participação das organizações é abaixo de 1%. O NA tem engajamento maior de sua população alvo, i.e. masculina, e nos EUA a participação dos gêneros no

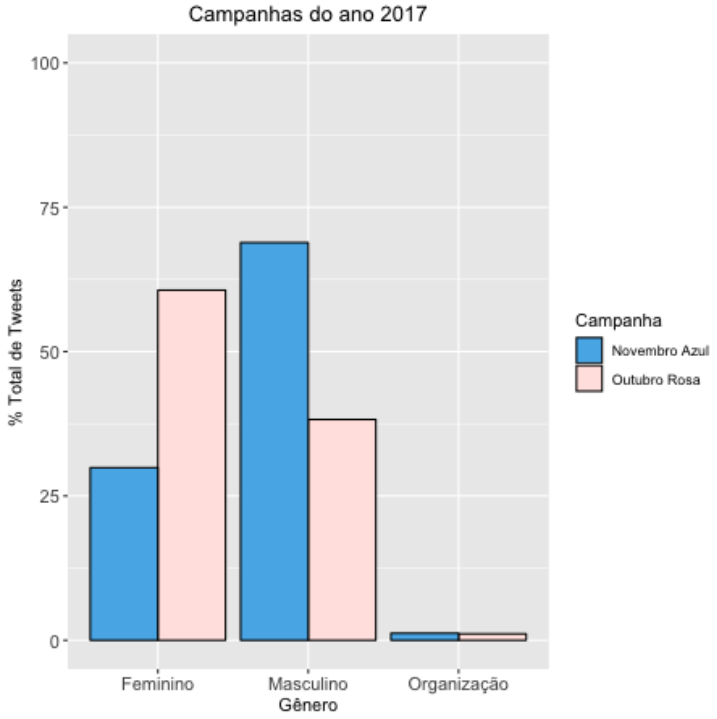


Figura 5.1: Distribuição por Gênero X Campanha do ano de 2017

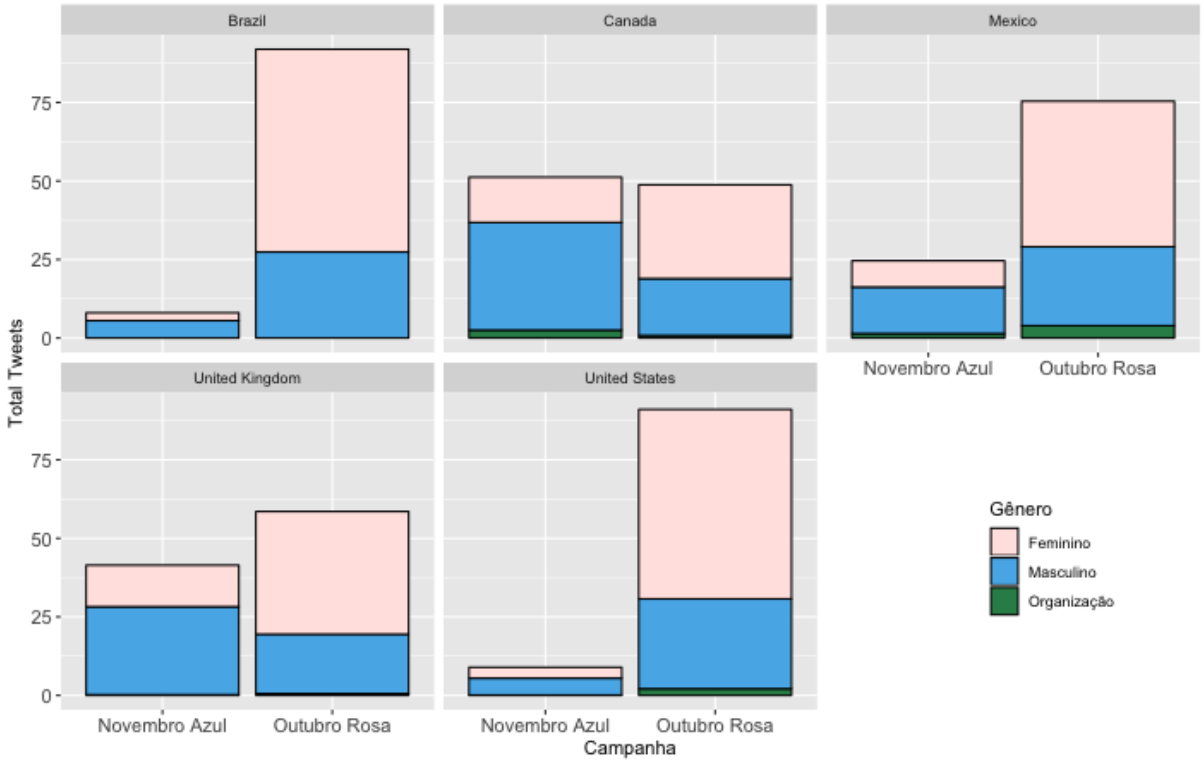


Figura 5.2: Distribuição de gênero por país das campanhas de 2017

NA é equilibrada. Já para o OR observam-se comportamentos mais consolidados em relação a seu público alvo, pois a participação feminina no Outubro Rosa é maioria em todos os países.

No que diz respeito à participação das faixas etárias, os usuários do tipo *Indivíduo* e *Celebridade* foram separados em um grupo com até 40 anos (-40), e acima de 40 anos (41+). Esses

grupos foram criados para verificar se as pessoas com mais de 40 anos, que são o público alvo da realização dos exames preventivos, possuem participação similar entre as campanhas do OR e NA, e os *tweets* de propósito geral. A Figura 5.3 apresenta a participação destes grupos etários nas duas campanhas. A participação de cada grupo nas duas campanhas é muito próxima, sendo que no OR o grupo 41+ tem uma participação levemente maior ao NA. Para estabelecer um *baseline*, indicamos pela linha horizontal vermelha, o índice de participação destas mesmas faixas etárias na postagem de *tweets* de propósito geral. Nestes *tweets*, a participação é de 95.2% e 4.8% para os grupos -40 e 41+, respectivamente (SLOAN et al., 2015). O alto índice de participação do grupo 41+ encontrado nos *tweets* deste trabalho, comparado a *tweets* em geral, indica que ambas campanhas estão atingindo os seus respectivos públicos alvo.

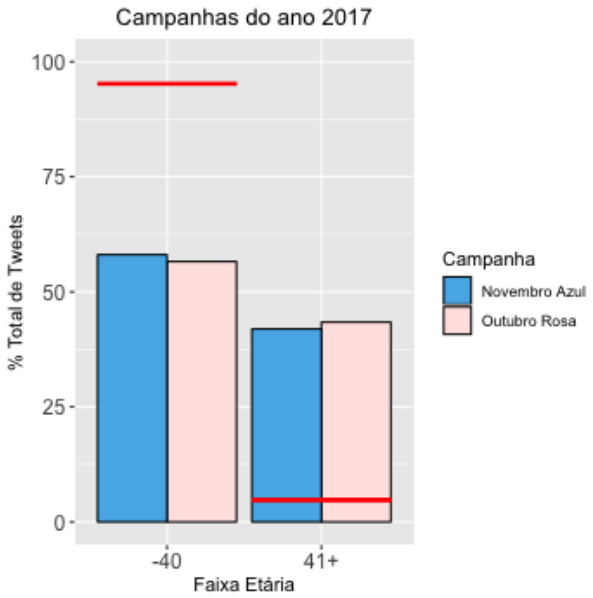


Figura 5.3: Distribuição por Faixa Etária X Campanha do ano de 2017

Na avaliação da distribuição da faixa etária por país (Figura 5.4), com exceção do Brasil e México, o nível de participação do público alvo (faixa etária de 41+) é maior do que no padrão da participação deste grupo nos *tweets* de propósito geral (4.8% (SLOAN et al., 2015)). No Brasil há maior participação do público que não é alvo da campanha (-40), enquanto que no México há um equilíbrio entre as duas populações.

Conclui-se assim que ambas as campanhas atingem seu público alvo, no tocante à gênero e faixa etária, mas que os níveis de consciência e engajamento são afetados pela cultura ou políticas próprias a cada país. A participação em geral das organizações em ambas as campanhas é bastante limitada.

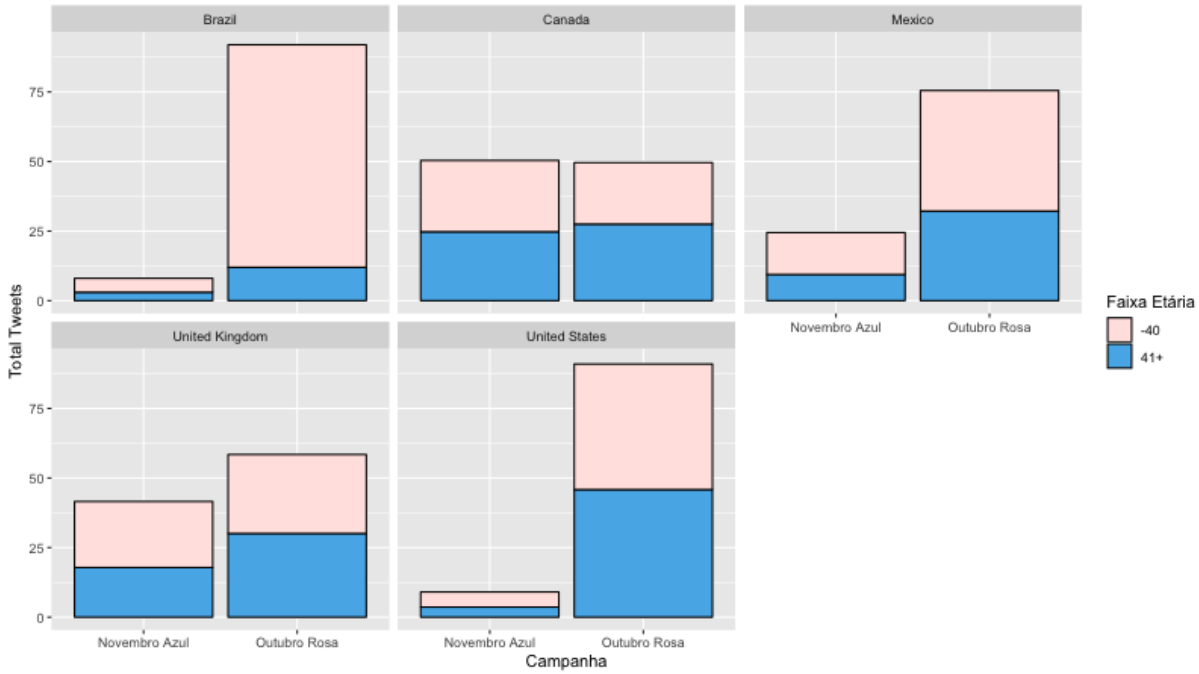


Figura 5.4: Distribuição de faixa etária por país das campanhas de 2017

## 5.2 QP 2: As campanhas apresentam características temporais similares?

Nesta seção realizamos uma avaliação de atividades temporais durante as campanhas, averiguando se existem diferenças e semelhanças entre as campanhas nos diferentes países considerados.

A distribuição de *tweets* por data de postagem são apresentados na Figura 5.5 utilizando o total de *tweets* coletados referentes ao ano de 2017. Verifica-se que a quantidade de *tweets* nos dias que precedem a cada campanhas é relativamente baixa, quando comparada comparada ao volume no período oficial de cada campanha (início no marco 0 do eixo $X$). As duas campanhas apresentaram um pico no início de seu respectivo mês. No caso do NA, após os primeiros dias a participação observa-se um declínio seguido de estabilização. A campanha do OR, por outro lado, teve três (3) picos (um deles no início da campanha), sendo o declínio da participação observado somente ao final do mês.

A Figura 5.6 detalha esta atividade para cada país. Canadá, Reino Unido e Estados Unidos revelam similaridade nas suas atividades detalhadas por campanha, ainda que apresentem proporções de participação total bastante diversos em termos de volume de engajamento (Seção 5.1). Para o NA, observa-se um pico de *tweets* nos primeiros dias da campanha, seguida de uma posterior estabilização, enquanto no OR existe uma estabilidade ao longo da campanha. Já o Brasil e México apresentam comportamentos distintos quando comparados a estes três países.

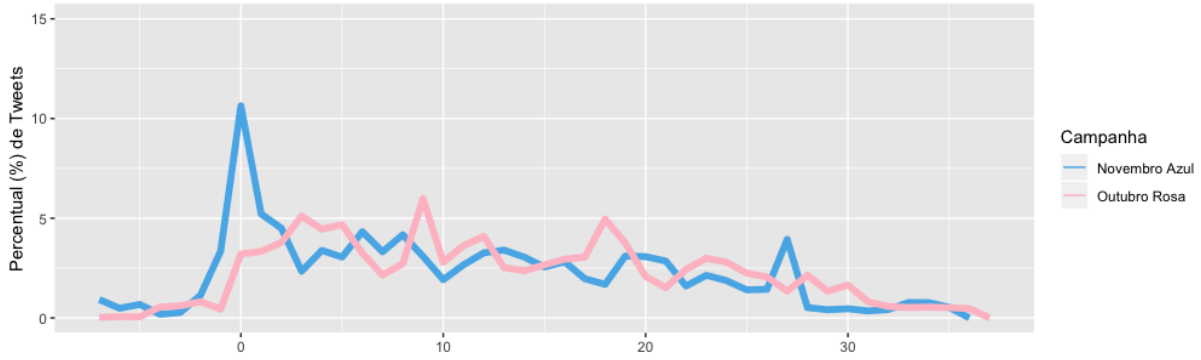


Figura 5.5: Distribuição por dia da campanha dos tweets das campanhas de 2017

O Brasil apresenta um grande pico no início do NA, bem maior que o pico inicial da campanha OR, seguido de bastante instabilidade de participação ao longo de ambas campanhas. O México também apresenta bastante instabilidade de atividade durante os períodos, mas em um padrão distinto do Brasil.

Para confirmar estas semelhanças e diferenças entre os países, calculamos as correlações entre as séries do OR e NA de todos pares de países pela medida de Normalized Cross-Correlation. Os resultados são apresentados na Tabela 5.1 para o OR e NA nas áreas rosa e azul, respectivamente. É possível averiguar que de fato o Canadá, Reino Unido e Estados Unidos apresentam similaridade entre as atividades temporais das campanhas, com correlações acima de 0.90 no NA e OR. Também o México, que aparentemente apresenta mais irregularidades ao longo do mês, indica uma similaridade maior que 0.90 com estes 3 países. Em termos das diferenças, pode-se confirmar também que o Brasil não possui correlações fortes com os demais países.

A Figura 5.7 apresenta o percentual acumulado dos *tweets* ao longo dos dias da campanha e evidencia a maior atividade do NA no início da campanha identificada anteriormente. Pode-se verificar que no 10º dia o NA já atinge 28.45% dos *tweets* postados, enquanto que a

Tabela 5.1: Correlação do nível de atividade por campanha do ano de 2017. As correlações OR e NA estão representadas nas cores rosa e azul.

| | Brazil | United States | United Kingdom | Canada | Mexico |
|---|---|---|---|---|---|
| Brazil | | 0.386 | 0.354 | 0.378 | 0.628 |
| United States | 0.786 | | 0.927 | 0.942 | 0.914 |
| United Kingdom | 0.718 | 0.972 | | 0.975 | 0.909 |
| Canada | 0.721 | 0.968 | 0.978 | | 0.924 |
| Mexico | 0.717 | 0.926 | 0.933 | 0.946 | |

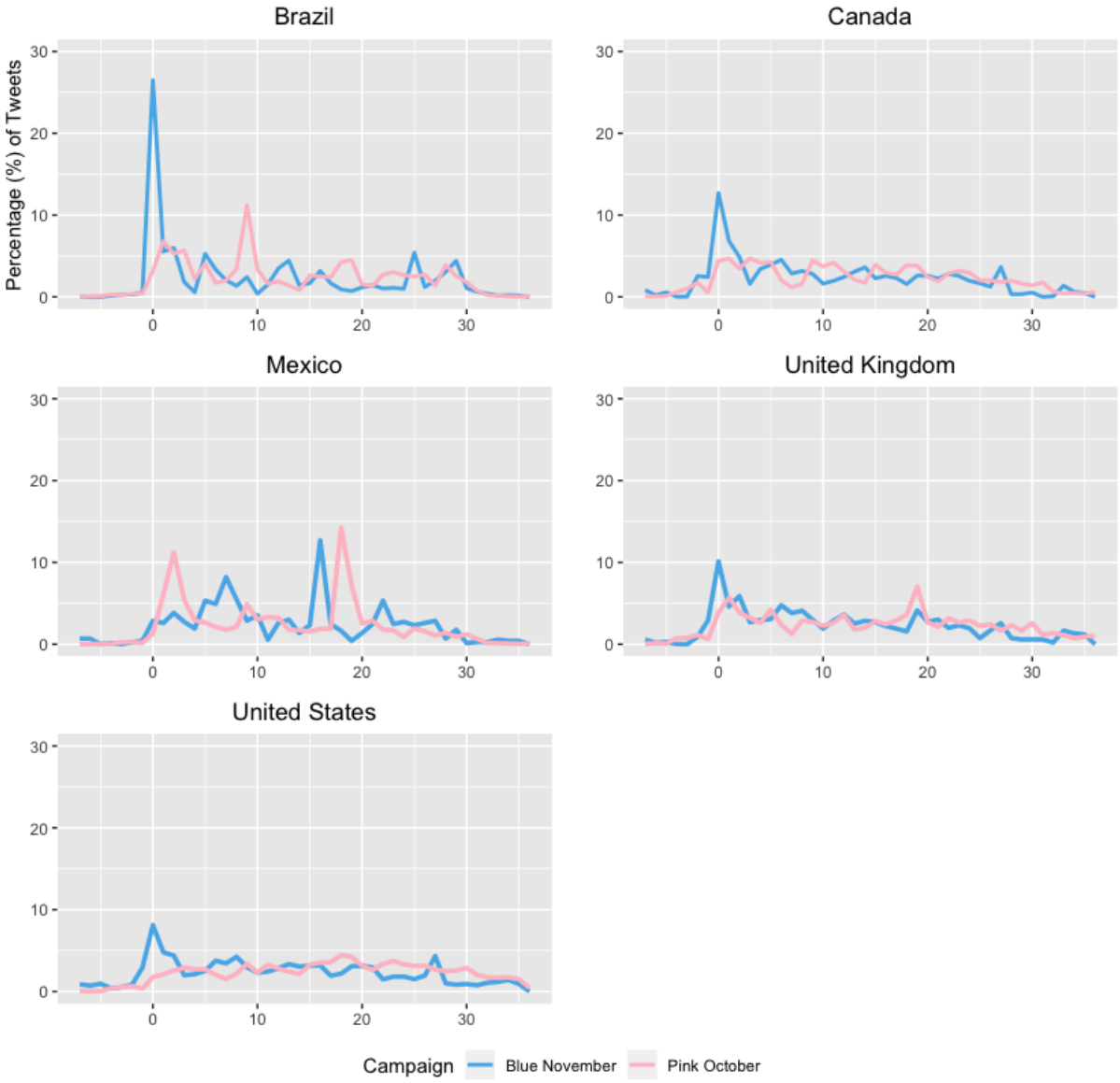


Figura 5.6: Distribuição dos tweets da campanha de 2017 por país

campanha do OR atinge aproximadamente esse percentual somente no 16º dia da sua campanha. No entanto, as duas curvas apresentam regularidade no seu crescimento durante o mês. As exceções desse crescimento regular são esse crescimento acelerado já mencionado do NA no início do seu mês, um declínio do NA mais acentuado no final, e um crescimento do OR por volta do 25º dia.

Concluímos que as campanhas do OR e NA têm padrões de atividades levemente distintos. No NA as atividades estão concentradas no início da campanha, enquanto no OR, picos adicionais demonstram um esforço de manter a campanha ativa ao longo de todo o período. Estes padrões são observados em diferentes países para ambas as campanhas, sendo Canadá, Reino Unido e Estados Unidos os países com maior similaridade entre os padrões de atividade.

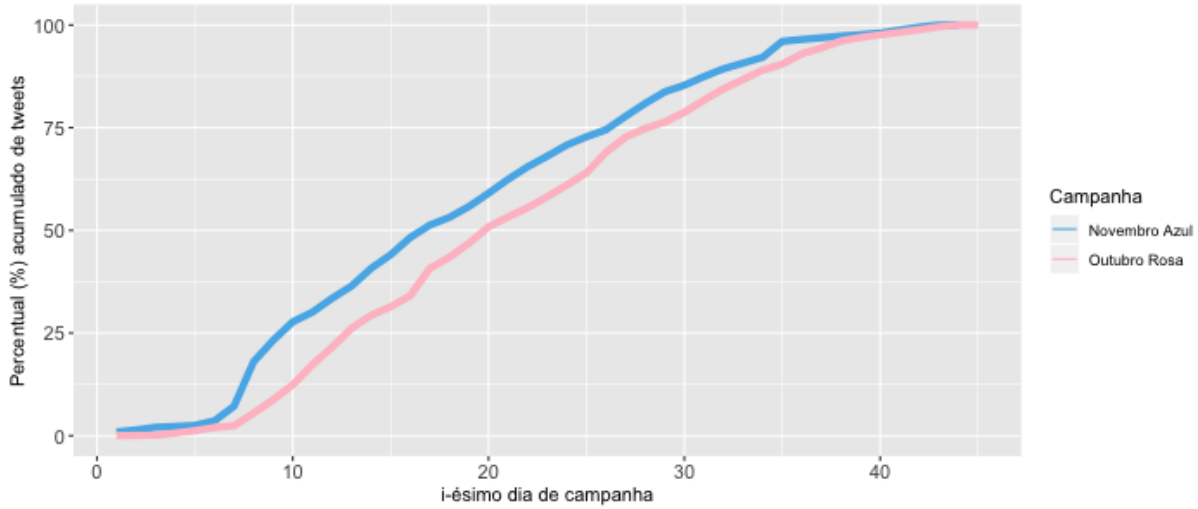


Figura 5.7: Percentual Acumulado (%) dos Tweets por Dia de Campanha em 2017

### 5.3 QP 3: As campanhas apresentam abrangência similar na rede social?

Realizamos uma avaliação de abrangência para saber como os *tweets* atingem as pessoas nas campanhas. A Tabela 5.2 mostra o número de *tweets* de cada campanha em 2017, o número de usuários envolvidos, bem como a média de *tweets* por usuário. A tabela detalha estes números por tipo de usuário. Nota-se que os números de usuários e *tweets* são muito superiores para o OR em todas as categorias. Com exceção das celebridades, a média de *tweets* por usuário também é maior no OR, indicando que os participantes do OR fazem mais postagens do que os envolvidos no NA.

Tabela 5.2: Características dos *tweets* por tipo de usuário em 2017

| | **Organizações** | | **Indivíduos** | | **Celebridades** | | **Total** | |
|---|---|---|---|---|---|---|---|---|
| | OR | NA | OR | NA | OR | NA | OR | NA |
| Nº usuários | 336 | 98 | 103308 | 24951 | 131 | 52 | 103775 | 25101 |
| Nº tweets | 1185 | 228 | 224142 | 40874 | 201 | 82 | 225528 | 41184 |
| **Média tweets** | **3.52** | **2.32** | **2.16** | **1.63** | **1.53** | **1.57** | **2.17** | **1.64** |

Sabendo que os *tweets* podem ser retuitados e atingir uma rede maior do que apenas os seguidores imediatos do perfil que postou a mensagem, buscamos compreender as relações por meio dos *retweets* entre os usuários. Para cada campanha, construímos um grafo que contém a estrutura das relações através *retweets* entre os usuários. Cada vértice do grafo representa um usuário, e as arestas que os conectam representam as conexões entre os usuários via *retweets*. O gráfico relativo ao OR é mostrado na Figura 5.8, e ao NA, na Figura 5.9. As cores dos vértices representam um tipo de estrutura (i.e. subgrafo), o que permite diferenciar os agrupamentos

dentro do grafo. É possível observar que o grafo do Outubro Rosa é mais denso que o grafo do Novembro Azul. Isso reflete as estruturas envolvidas, onde, de forma geral, as estruturas do Outubro Rosa possuem mais vértices (usuários) a serem representados no grafo. Por exemplo, a estrutura central no gráfico do Outubro Rosa (Figura 5.8) envolve 82520 vértices (usuários), enquanto que a a estrutura central do Novembro Azul (Figura 5.9) representa 12350 usuários, ou seja, apenas 14.9% do número do Outubro Rosa.

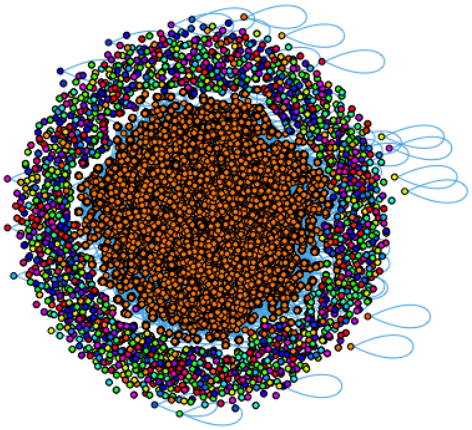

Figura 5.8: Conexões via *retweets* entre usuários do OR em 2017

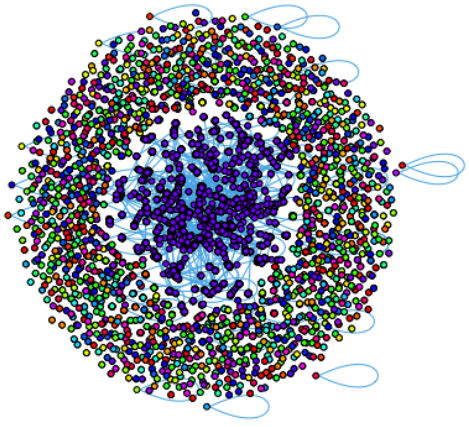

Figura 5.9: Conexões via *retweets* entre usuários do NA em 2017

Algumas dessas estruturas (subgrafos) são detalhadas nas Figuras 5.10 e 5.11 para o OR e NA, respectivamente. Estas figuras mostram, para cada subgrafos das redes de conexão entre os usuários, o número de usuários (V), conexões entre os usuários através de *retweets* (E), e o número de vezes que esse par de informações ocorre (N), ordenados nas linhas de cada figura pelo número de usuários em cada estrutura. Faremos referência a estes subgrafos na forma de linhas e colunas (e.g. [1,1] para o subgrafo na primeira linha e coluna). Vemos as estruturas

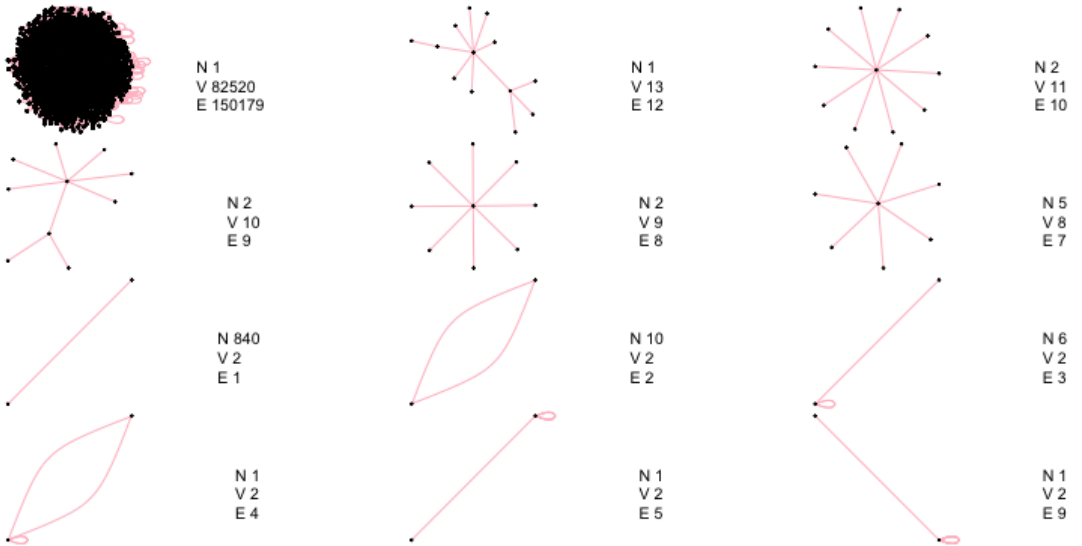


Figura 5.10: Estruturas de retweets entre usuários do OR que se retuítam (2017)

centrais dos grafos nas figuras 5.8 e 5.9 na posição [1,1] das matrizes de estruturas nas Figuras 5.10 e 5.11, respetivamente.

Estas figuras mostram alguns padrões interessantes. É possível observar o alto número de usuários centralizados que são retuitados. Para o OR, podemos observar isto nas estruturas [1, 1], [1,2] e [1,3] da Figura 5.10, repetindo-se nas linhas seguintes com menor alcance. Para o NA este padrão também é observado (e.g. [1,1], [1,2] e [1, 3]). Outro padrão interessante é a profundidade que os *tweets* atingem na rede, onde existem vários níveis de conexões entre os usuários. Por exemplo, para o OR a estrutura [1, 2] atinge 5 níveis, e para o NA [2, 1] até 9 níveis.

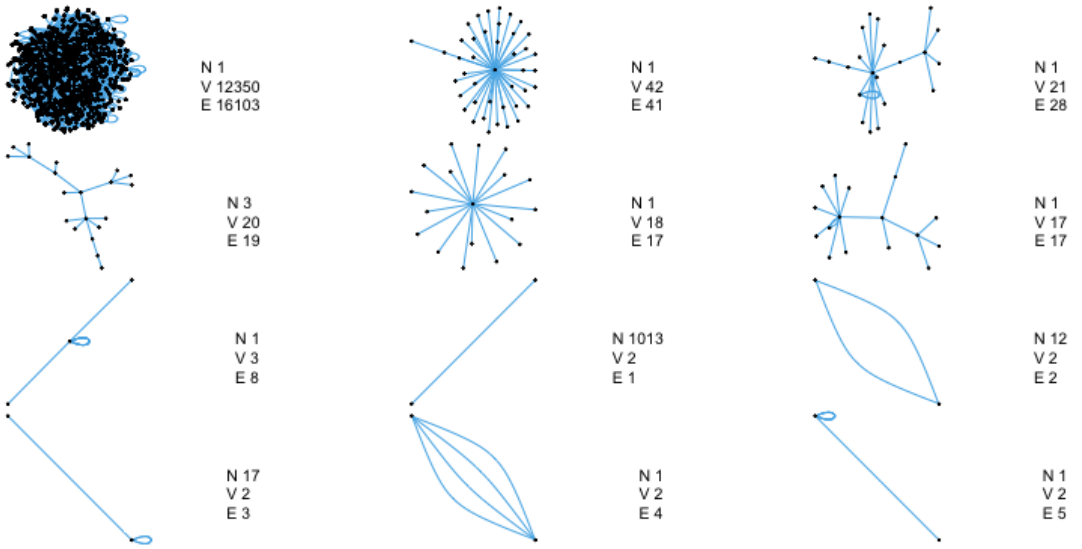


Figura 5.11: Estruturas de retweets entre usuários do NA que se retuítam (2017)

A Figura 5.12 apresenta um diagrama de caixa (*boxplot*) para avaliação dos *retweets* de cada campanha. Podemos observar que a mediana das duas campanhas está definida no valor 2, *i.e.*, metade dos *tweets* retuitados são retuitados até duas vezes. Na evolução do diagrama identificamos que o Outubro Rosa possui maior abrangência com 5 e 11 *retweets* no terceiro quartil e valor máximo, contra 4 e 8 no Novembro Azul.

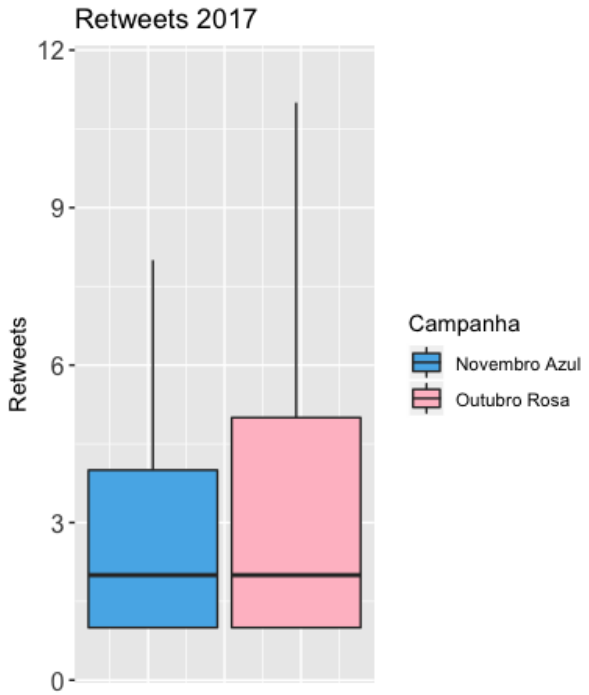


Figura 5.12: Distribuição da frequência de *retweets* por campanha

Na Figura 5.13 fazemos essa avaliação por país. Podemos identificar que o padrão geral (Figura 5.12) repete-se no México, Reino Unido e Estados Unidos. Observa-se que a mediana de *retweets* para o OR na Inglaterra é levemente superior (3). O Canadá apresenta comportamentos exatamente iguais nas duas campanhas, com 4 e 8 *retweets* no terceiro quartil e valor máximo, respectivamente. Já o Brasil apresenta um padrão similar no tocante à proporção entre as campanhas, mas os valores das medianas e demais quartis são mais baixos. No Novembro Azul, o primeiro quartil e mediana são iguais (apenas 1 *retweet*), assim como o terceiro e quarto quartis (2 *retweets*), ou seja, metade dos *tweets* são retuitados até uma vez, a quase 100% são retuitados no máximo duas vezes. O Outubro Rosa apresenta um panorama um pouco mais ativo, com o terceiro e quarto quartis indicando 3 e 7 *retweets*, respectivamente.

As Figuras 5.14 e 5.15 detalham o número de *retweets* por data de criação do *post* original para as campanhas do Outubro Rosa e Novembro Azul, respectivamente. Verifica-se pela mediana de cada campanha, indicadas pelas linhas na horizontal, que a campanha do

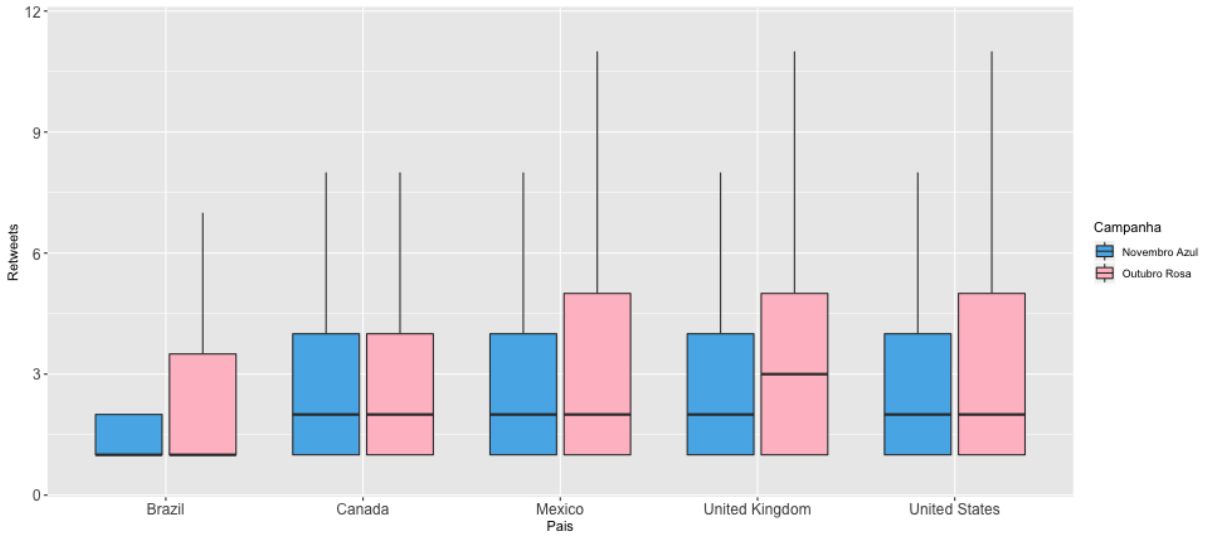


Figura 5.13: Frequência de *retweets* por campanha

Outubro Rosa (12) possui republicações de *tweets* mais altas ao Novembro Azul (9). Ou seja, os *tweets* do Outubro Rosa atingem números mais altos de *retweets* do que o Novembro Azul. Isso indica maiores níveis de atividades e propagação de *tweets* no Outubro Rosa.

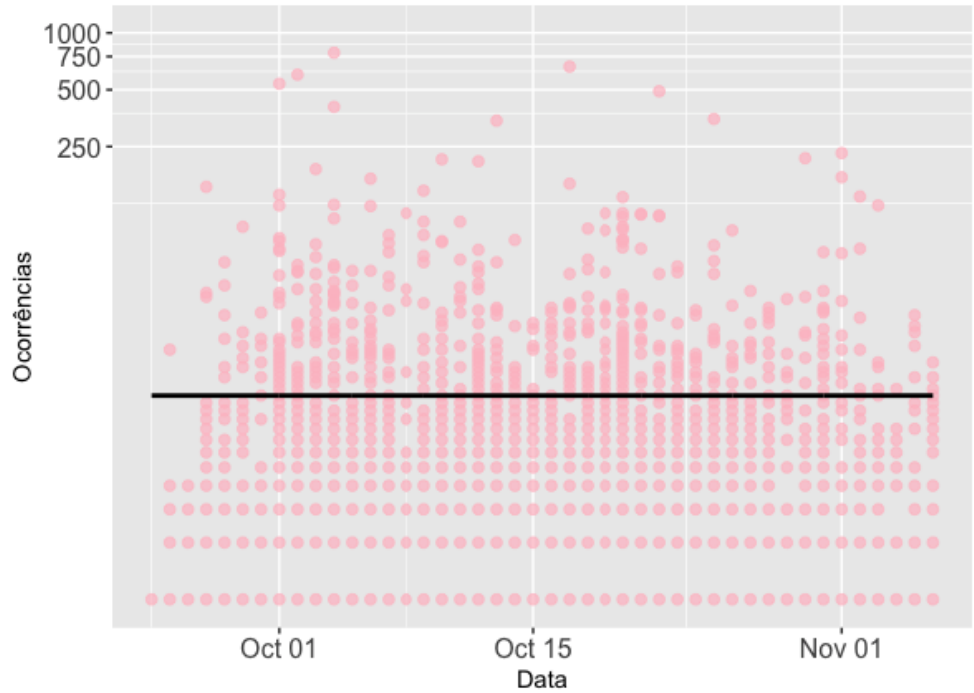


Figura 5.14: Frequência de *retweets* por data do Outubro Rosa

Finalmente, a distribuição do grau de conexões por tipo de usuário a partir de *retweets* é exibida na Figura 5.16. O eixo $x$ da figura representa o número de vezes em que um *tweet* foi republicado, e o eixo $y$ representa o número de vezes em que este número de republicações ocorreu. A maioria dos usuários são retuitados poucas vezes, e pouquíssimos usuários

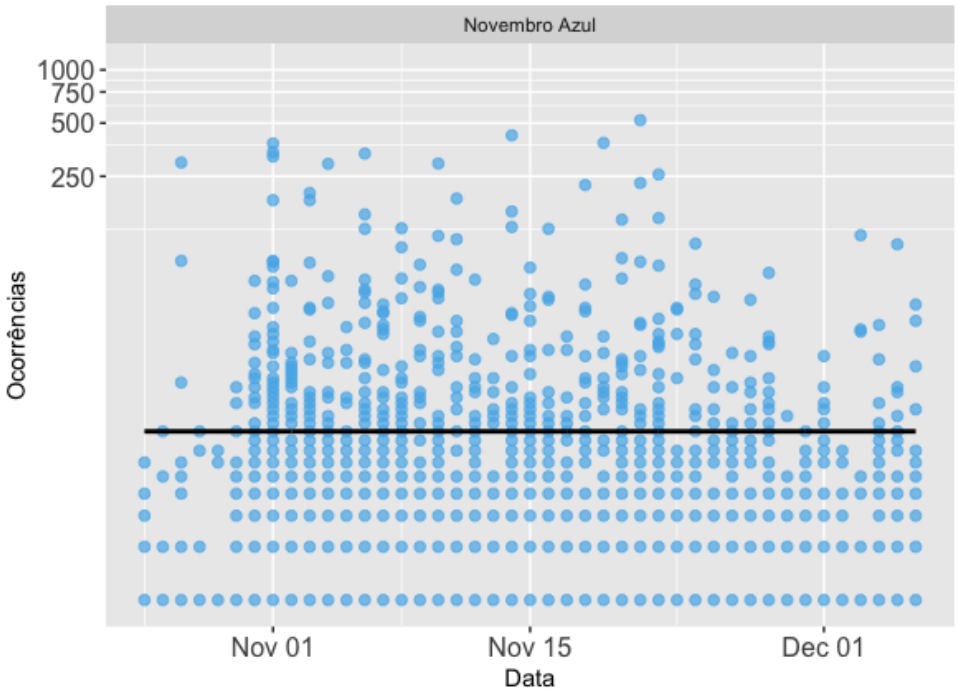


Figura 5.15: Frequência de *retweets* por data do Novembro Azul

possuem várias conexões por *retweets*, indicando que a maioria dos usuários não tem influência na propagação dos *tweets* na rede do Twitter. Ainda, na categoria Indivíduo, observa-se que o OR apresentou uma maior capacidade de propagação de informação comparado ao NA. Nas categorias Organização e Celebridade as duas apresentam equilíbrio. Pode-se concluir que a abrangência de propagação dos *tweets* originais através de *retweets* acaba sendo maior no Outubro Rosa. Isso pelo fato da categoria Indivíduos possuir graus de *retweets* do OR mais frequentes do que o NA, e os *tweets* atingirem graus de *retweets* que o NA não atinge.

Conclui-se que as campanhas do OR e NA no Twitter não apresentam a mesma abrangência. Além de a campanha do OR engajar um número de usuários muito maior, estes usuários são em média muito mais ativos, pois postam em média 32 pontos percentuais a mais de *tweets* (2.16 *tweets* no OR vs. 1.63 *tweets* no NA). O alcance destes *tweets* também é bem maior na campanha do OR, já que o número de *retweets* também é comparativamente maior. Essas conclusões se aplicam a categoria de usuários classificados como Indivíduos, cuja a influência é mais forte no OR. Já as Celebridades e Organizações possuem participação muito equilibrada comparando-se as duas campanhas. Finalmente, observa-se que celebridades e organizações ainda têm pouco destaque na divulgação de ambas as campanhas.

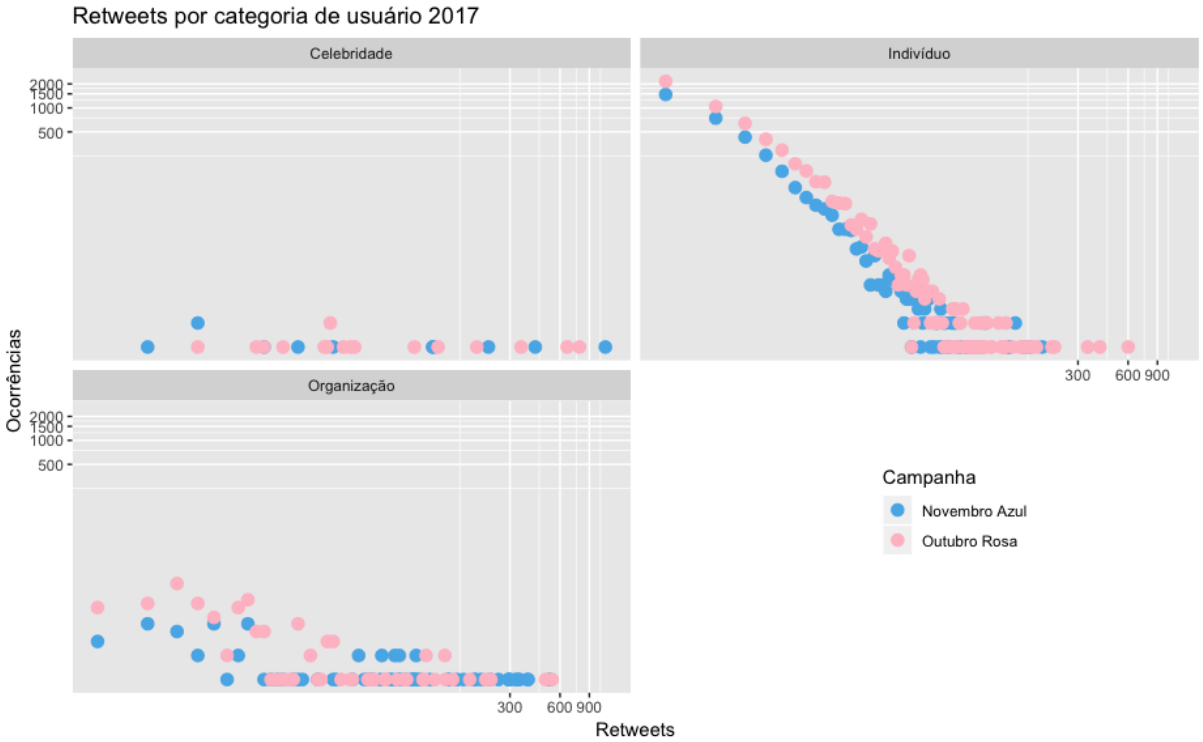


Figura 5.16: Distribuição de grau das conexões entre os usuários (log x log)

## 5.4 QP 4: As campanhas abordam tópicos similares de conteúdo?

Nesta seção são apresentadas as análises referentes aos tópicos dos *tweets* postados durante as campanhas. O tópico de cada *tweet* foi categorizado automaticamente conforme descrito na Seção 4.4. Como já mencionado, a análise dessa questão de pesquisa foi limitada para *tweets* do idioma inglês, devido à extensão das análises contemplando todos os idiomas.

A distribuição das categorias de tópicos para as campanhas do OR e NA do ano de 2017 são apresentadas nas Figuras 5.17 e 5.18, respectivamente. Os tópicos mais discutidos nas campanhas foram *Fundraising*, *Early Detection/Diagnosis*, *Loved Ones* e *Wear Pink/Blue*. Contudo, existem diferenças significativas entre as campanhas e países.

Pode-se observar que a campanha do Outubro Rosa apresentou uma maior diversidade nos assuntos postados, sendo as categorias mais frequentes *Fundraising* (31%), *Loved Ones* (30%) e *Early Detection/Diagnosis* (17%). A quantidade de postagens relacionados às categorias *Wear Pink* (4%) e *Treatments* (6%) foi significativamente menor, e não foi encontrado nenhum tópico relacionado às categorias *Resentment* e *Walk & Runs / Sports*. Nota-se assim uma pequena mudança no ano de 2017, se comparado ao ano de 2012, quando as categorias descritas na Tabela 3.1 foram identificadas (THACKERAY et al., 2013). Ressalta-se que não foi possível categorizar 12% dos *tweets*, ou seja, *tweets* cujos tópicos não apresentaram similaridade com nenhuma categoria do referencial, sendo estes então rotulados como *Others*.

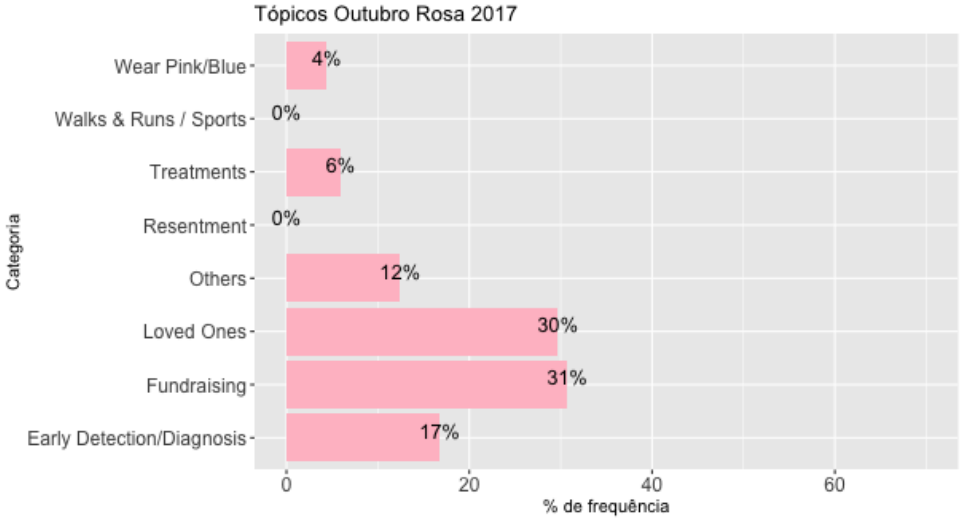


Figura 5.17: Resumo de tópicos do Outubro Rosa 2017

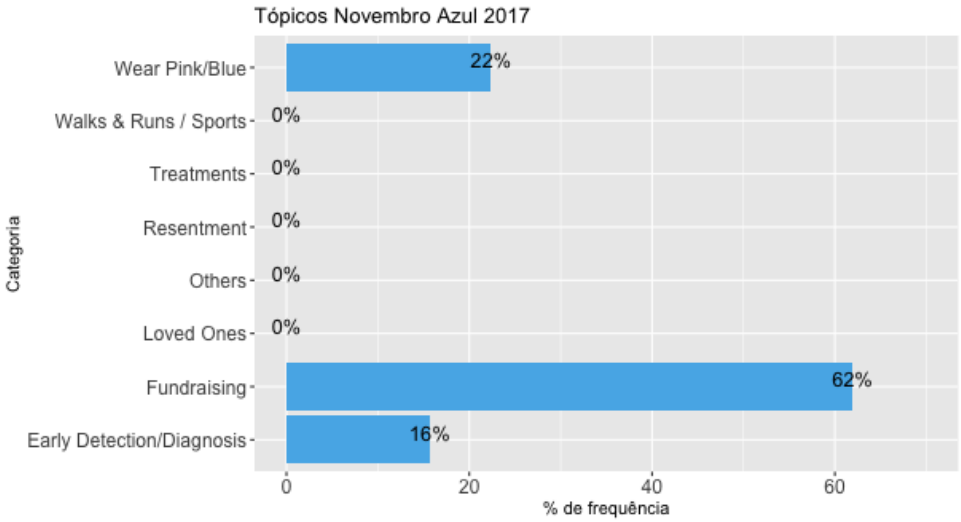


Figura 5.18: Resumo de tópicos do Novembro Azul 2017

Já em relação às postagens da campanha masculina NA, os assuntos são muito mais limitados (Figura 5.18). A categoria *Fundraising* concentra 62% dos *tweets*, seguida das categorias *Wear Blue* (22%) e *Early Detection/Diagnosis* (16%). Todos os *tweets* referentes à campanha do NA em 2017 foram categorizados.

Assim, percebemos diferenças nas postagem de *tweets* nas duas campanhas, o que talvez possa ser atribuído à maturidade das mesmas. Embora a campanha do Outubro Rosa aborde o levantamento de recursos (*Fundraising*) e a detecção do câncer de forma prematura (*Early Detection/Diagnosis*), ela também ressalta a importância de cuidados com os entes queridos e relatos pessoais (*Loved Ones*). No Novembro Azul isso não acontece, já que o foco está mais

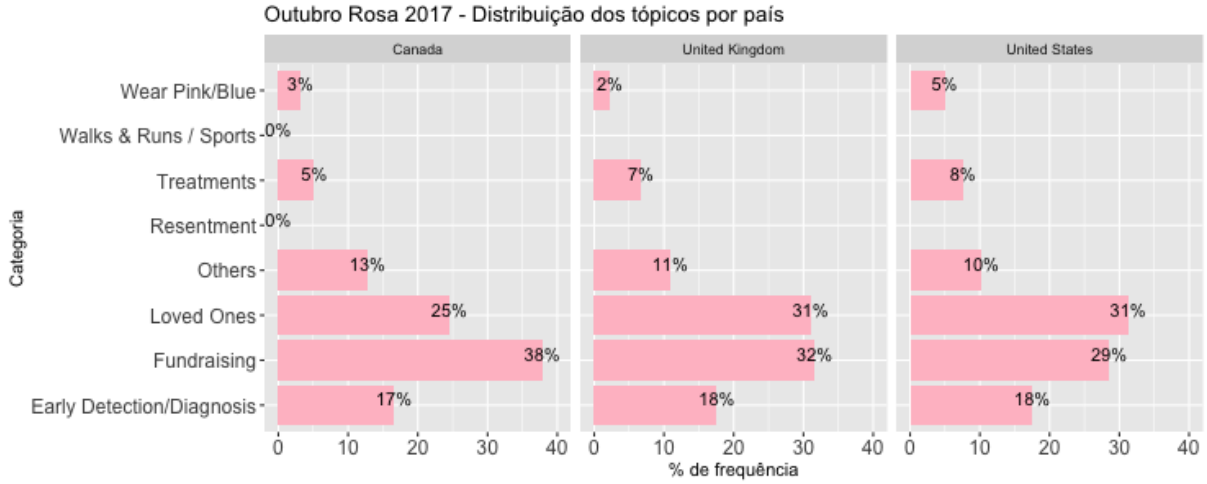


Figura 5.19: Distribuição dos tópicos do Outubro Rosa 2017 por país

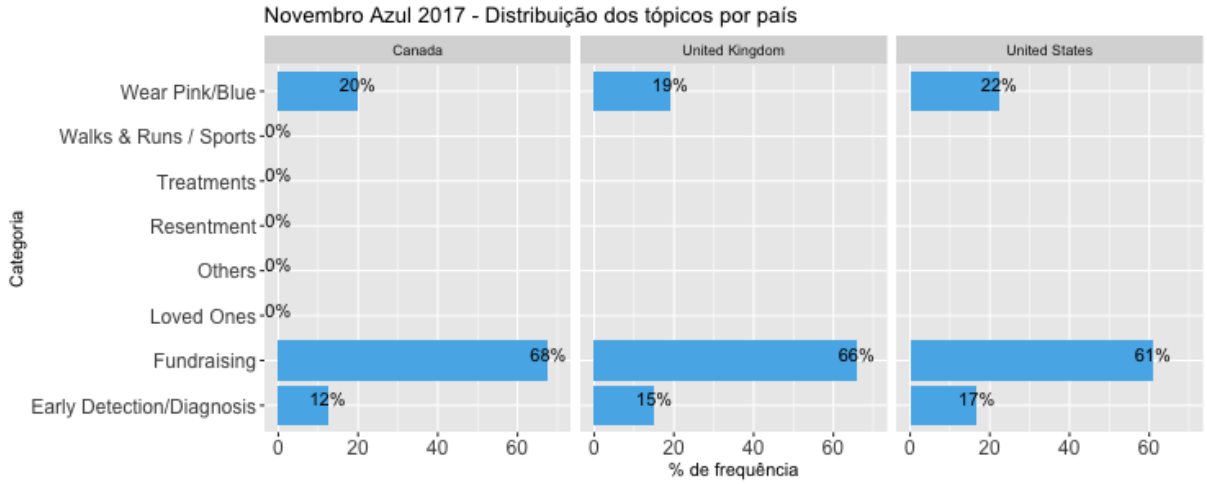


Figura 5.20: Distribuição dos tópicos do Novembro Azul 2017 por país

limitado a levantar recursos financeiros (*Fundraising*) e engajar as pessoas na campanha (*Wear Blue*).

Ao detalharmos essas informações por país, percebemos realidades muito similares à distribuição geral identificada nas Figuras 5.17 e 5.18. No caso do Outubro Rosa (Figura 5.19), apenas o Canadá apresenta alterações significativas quanto às categorias *Loved Ones* e *Fundraising*. A categoria *Loved Ones* recebe menos atenção (5 pontos percentuais - pp) se comparada à distribuição geral. Por outro lado, a categoria *Fundraising* representa 8 pp a mais quando comparada ao padrão geral.

Para o Novembro Azul (Figura 5.20), a categoria *Fundraising* predomina nos 3 países, principalmente no Canadá e Reino Unido com 68% e 66%, respectivamente. Nos EUA esta categoria também possui alta representatividade (61%), mas outros assuntos têm também destaque: *Wear Blue* (22%) e *Early Detection/Diagnosis* (17%).

A distribuição das categorias ao longo do mês de cada campanha é apresentada nas Figuras 5.21 e 5.23. As linhas nos gráficos representam as categorias da campanha. O eixo $x$ representa o $i$-ésimo dia de campanha, e o eixo $y$ indica o percentual de participação da categoria do tópico no dia $x$ em relação ao total de *tweets* analisados. Na Figura 5.21 identificamos que a campanha começou com uma ênfase emocional (Loved Ones), sendo que por volta do 9º dia de campanha o tema *Fundraising* passou a ser o foco principal. A ênfase em diagnóstico/prevenção existe sempre ao longo da campanha, com alguns picos específicos. O direcionamento para a categoria *Weak Pink* com foco na identificação da campanha ocorre somente no final do período. Quando detalhamos essa distribuição por país na Figura 5.24, percebemos que o Canadá e Reino Unido apresentam comportamentos similares ao da Figura 5.21, com vários picos de *tweets* e bastante variação, enquanto que os EUA apresentam um equilíbrio maior ao longo do mês, e apenas uma maior elevação dos níveis da categoria *Loved Ones* em alguns momentos.

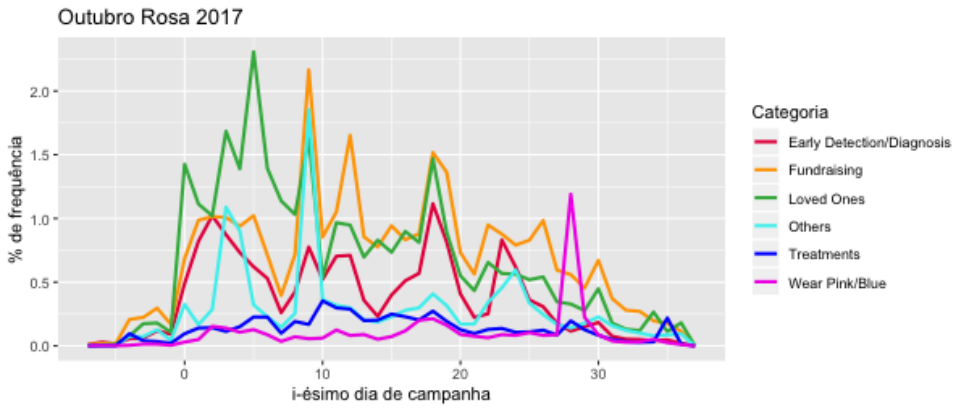


Figura 5.21: Distribuição dos tópicos ao longo da campanha Outubro Rosa 2017

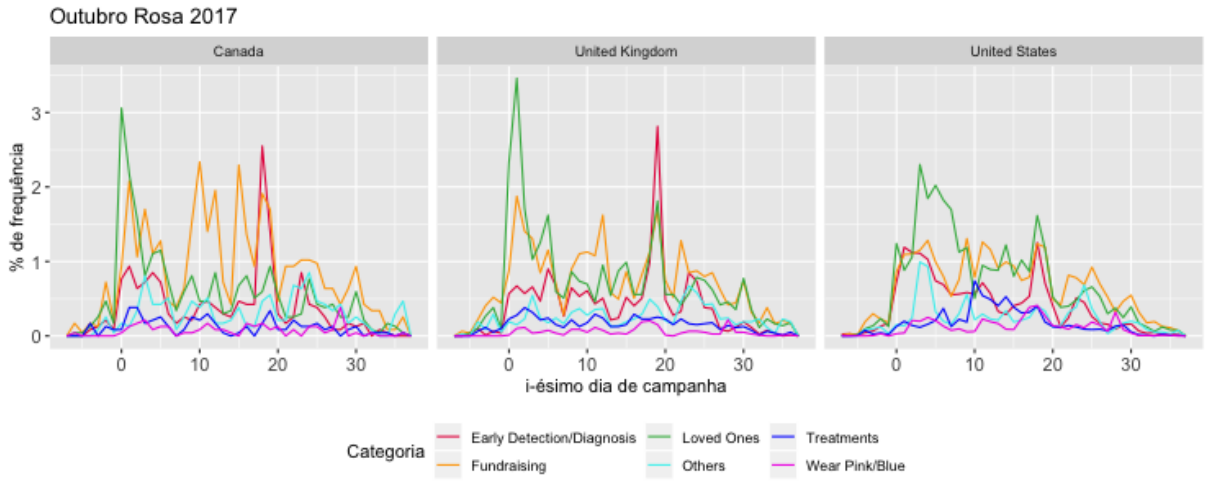


Figura 5.22: Distribuição dos tópicos ao longo da campanha Outubro Rosa 2017 por país

O Novembro Azul teve apenas 3 categorias de assuntos postados ao longo da campanha, e estas categorias apresentaram equilíbrio no comportamento ao longo do mês. Como podemos identificar na Figura 5.23, a categoria *Fundraising* teve uma prevalência maior ao longo de toda campanha. Apenas no início e fim a campanha apresentou equilíbrio em relação às demais categorias. Na Figura 5.24, percebemos que padrões muito similares de comportamento ao longo do mês são mantidos na distribuição por país, onde no 11º dia houve uma alteração mais significativa do tópico de *Early Detection/Diagnosis* no Reino Unido.

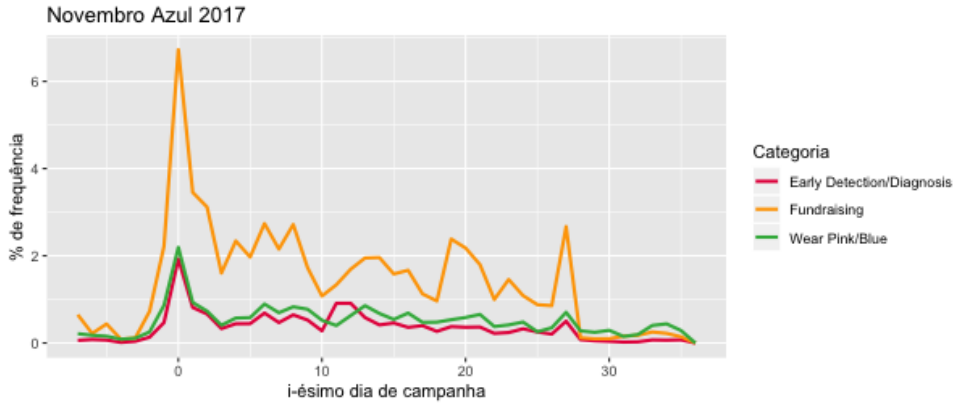


Figura 5.23: Distribuição dos tópicos ao longo da campanha Novembro Azul 2017

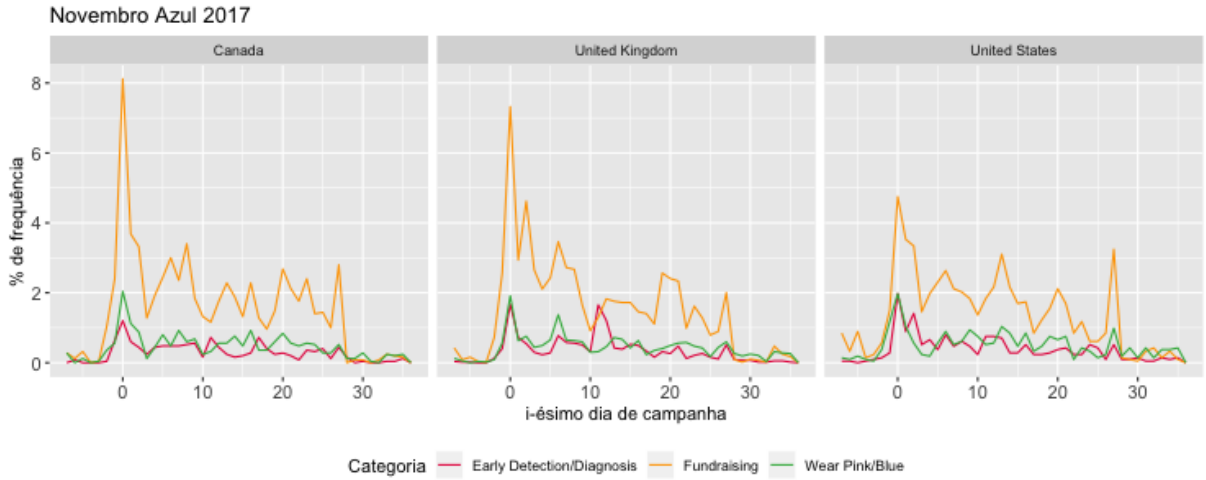


Figura 5.24: Distribuição dos tópicos ao longo da campanha Novembro Azul 2017 por país

Podemos concluir que *Fundraising* e *Early Detection/Diagnosis* são preocupações comuns dos *tweeters* em ambas as campanhas, já que estes dois assuntos são bastante divulgados. Já a categoria *Wear Pink/Blue* foi bastante abordada no Novembro Azul, recebendo bastante atenção nos 3 países avaliados. Contudo, ela não foi representativa no Outubro Rosa. A categoria *Treatments*, ainda que em percentual menor em relação aos demais assuntos, foi discutida

apenas no Outubro Rosa, assim como *Loved Ones*. Como a conscientização, a detecção precoce e o tratamento são o foco das duas campanhas, concluímos que elas atingiram parcialmente seu objetivo, e que estratégias são necessárias para motivar as pessoas a compartilharem mais suas preocupações, conscientização e experiências sobre câncer, principalmente o de próstata, e passarem pelo testes de triagem para melhorar a chance de sobrevivência. Uma hipótese a investigar que talvez explique estes resultados é que o Novembro Azul possa estar em um estágio de maturidade distinto do Outubro Rosa, buscando ainda promover a conscientização sobre o assunto.

A falta de tópicos categorizados em *Resentment* e *Walks and Runs / Sports* nos faz considerar que estas categorias propostas por Thackeray et al. (2013) devem ser reavaliadas para identificar se continuam pertinentes e utilizáveis nos dias atuais, ou se elas tiveram mudança significativa de vocabulário ao longo dos anos e devem sofrer adaptações. No que tange as demais categorias, dado a alta similaridade e adesão dos tópicos do LDA, consideramos que as mesmas continuam válidas e atuais. Já a categoria *Others* deve passar por uma etapa de avaliação mais profunda, a fim de verificar se palavras relacionadas a esta categoria podem dar lugar a novas categorias mais específicas.

## 5.5 QP 5: O comportamento das campanhas tem variado ao longo dos anos?

Nas seções anteriores, desenvolvemos diferentes análises centradas no comportamento das campanhas de 2017. Nesta seção analisamos se este comportamento variou em relação a anos anteriores (2014-2016) ou ao ano seguinte (2018).

### 5.5.1 Público Alvo

Na Seção 5.1 concluímos que, para o ano de 2017, as campanhas tiveram sucesso junto ao respectivo público alvo, considerando o gênero e a idade. Esta análise longitudinal tem por objetivo verificar se este padrão também foi observado nos demais anos analisados.

Na Figura 5.25 podemos avaliar a evolução do volume da participação dos gêneros masculino e feminino, e das organizações nas campanhas. A figura mostra que o padrão identificado no ano de 2017 também foi observado nos demais anos: o gênero que mais participa corresponde ao gênero alvo da campanha, *i.e.*, feminino no Outubro Rosa e masculino no Novembro Azul. Não há grandes variações na participação de organizações ao longo dos anos.


Figura 5.25: Volume de participação dos gêneros nas campanhas

Contudo, nesta visão longitudinal, também pudemos identificar uma grande diferença no volume de *posts* publicados em cada campanha ao longo dos anos. Considerando o ano do início da análise (2014), as duas campanhas tiveram períodos onde houve uma queda bastante significativa na participação (número absoluto de posts). As variações no volume são referentes a ambos os gêneros. Na campanha do Novembro Azul verificamos que o ano de 2014 foi o que apresentou maior maior volume de *posts*, mas que até 2017 o volume foi progressivamente caindo. Apenas o ano de 2018 apresentou uma recuperação em relação a anos anteriores, quando a campanha voltou ao nível de volume observado em 2015, mas ainda bastante longe do volume do ano de 2014. Já no caso do Outubro Rosa, visualizamos um comportamento um pouco diferente: a queda foi observada até o ano de 2016, mas em 2017 houve uma grande recuperação, sendo que o volume deste ano atingiu números que ainda não haviam sido alcançados. No ano de 2018 houve um novo aumento no volume de *tweets*, superando o ano de 2017.

Na Figura 5.26 avaliamos a evolução ao longo dos anos de participação do público alvo das campanhas, a saber, a faixa etária 41+, comparada a participação dessa faixa etária em *tweets* de propósito geral (linha constante vermelha em 4,8%). Utilizamos a proporção de *posts* sobre o volume total do ano, a fim de comparar os resultados, considerando a variação no volume acime destacada. Observa-se uma participação mínima de 32% (OR, 2014) e máxima de 43% (OR, 2017). Isto é bem superior ao *baseline* de postagem de *tweets* de propósito geral para esta faixa etária. Observamos uma tendência de crescimento desta participação até o ano

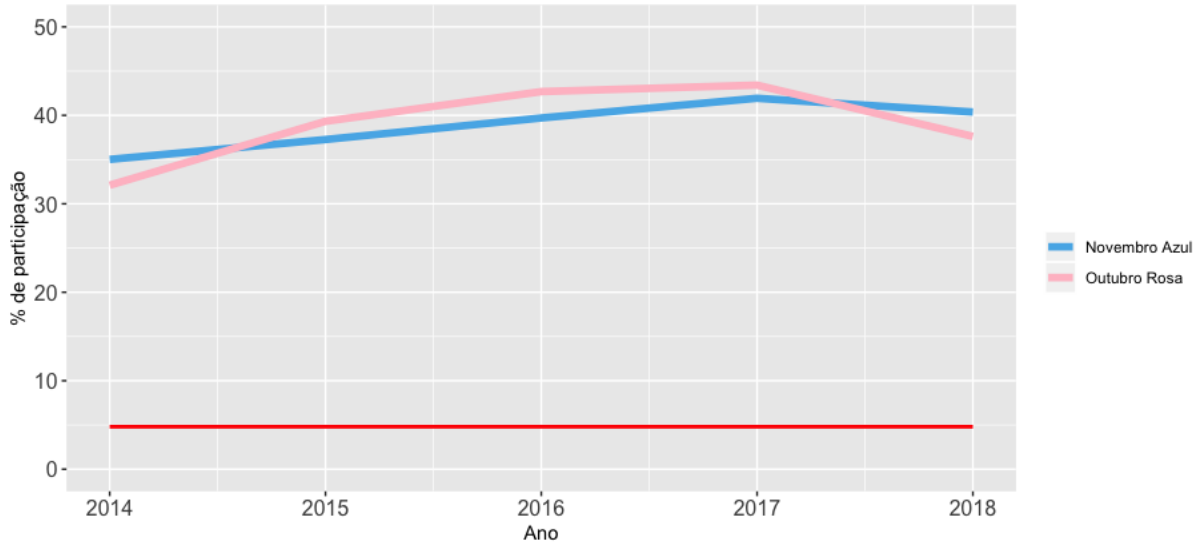


Figura 5.26: Porcentual do volume de participação da faixa 41+ nas campanhas

de 2017, sendo a do OR mais acentuada (11 pp entre 2014 e 2017 para o OR, contra 7 pp para o NA). Em 2018 houve uma queda de 6 pp e 2 pp nas campanhas OR e NA, respectivamente. Portanto, entendemos que, no quesito idade, as campanhas vêm atingido seu público alvo.

A Figura 5.27 apresenta o detalhamento da participação da faixa 41+ para os diferentes países. Ao avaliarmos estes gráficos, percebemos uma enorme variação na participação de usuários desta faixa etária, mas sempre acima do *baseline* de *tweets* de propósito geral. Para todos os países, exceto o Brasil a participação mínima no OR é de 35% (Estados Unidos - 2014) e máxima 55% (Canadá - 2017). Quanto ao NA, a participação mínima é 29% (México - 2015) e máxima 53% (Canadá - 2018). De uma forma geral, a tendência é de crescimento no OR, com leve queda no México no ano de 2017. Já para o NA, a tendência é de relativa estabilidade, com crescimento constante do Canadá.

Já a participação no Brasil é comparativamente bem menor em ambas as campanhas. No caso do OR, ela era crescente até 2017, quando teve uma queda significativa. Mas voltou a crescer no ano de 2018. No caso do NA ela é crescente ao longo dos anos, tendo alcançado o patamar de 37% nos anos de 2017/2018, equiparando-se a outros países (México, Estados Unidos e Reino Unido).

Concluímos que as campanhas vêm atingindo seu público alvo, mas com grande variação no volume de *posts* envolvidos nas campanhas. Dentre os países, o Brasil foi o que apresentou comportamento mais irregular e distinto dos demais países. Em relação ao Novembro Azul, o Brasil foi o que demonstrou maior crescimento junto a seu público alvo, assim como o Canadá.

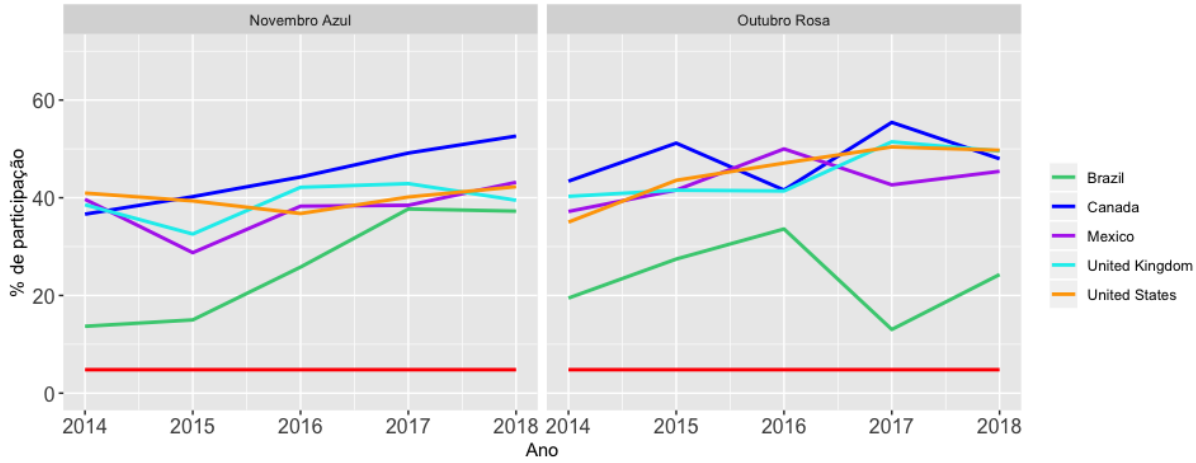


Figura 5.27: Volume de participação da faixa etária 41+ nos países

### 5.5.2 Propagação

Na Seção 5.2 caracterizamos a distribuição dos *tweets* por data de postagem para o ano de 2017. Observamos que de uma forma geral, o Outubro Rosa buscava manter a campanha ativa ao longo do mês, com alguns picos, enquanto que no Novembro Azul, a maior concentração de atividades era observada no início da campanha (Figura 5.5). Encontramos correlações fortes neste comportamento para os diferentes países, com exceção do Brasil.

Para analisar se este comportamento se repete ao longo dos anos para cada país, geramos uma distribuição geral por ano, que é a sumarização das atividades no Twitter por ano e campanha, mas indiferente do país. A partir disso, correlacionamos as distribuições de cada campanha/ano/país com a respectiva distribuição geral (campanha/ano) pela medida de Normalized Cross-Correlation. Assim, podemos avaliar se houve alterações de comportamentos das campanhas ao longo do respectivo mês nos diferentes países. A Figura 5.28 resume as medida de correlação dessas distribuições.

Podemos observar na Figura 5.28 que de modo geral as correlações são bastante fortes. Em todos os países e campanhas, exceto o Brasil, os índices são superiores a 0,9. Novamente o Brasil possui um comportamento levemente distinto ao longo dos anos, mas as correlações são também consideradas fortes, à exceção de um único caso. Mais especificamente, no Novembro Azul, a correlação varia entre 0,74 e 0,89, que são índices de correlação ainda fortes. No Outubro Rosa temos uma situação interessante, onde a correlação é acima de 0,9 nos anos 2014 a 2016, mas o ano de 2017 apresenta um valor de correlação muito menor, com valor 0,45 (fracamente correlacionado). No ano de 2018 a correlação de 0,71 revela uma relação mais próxima às atividades da base geral, mas ainda está num índice relativamente baixo se comparado aos outros países/campanhas.

| | | Ano | | | | |
|---|---|---|---|---|---|---|
| Pais | Campanha | 2014 | 2015 | 2016 | 2017 | 2018 |
| Brazil | Novembro Azul | 0.7620 | 0.8430 | 0.8770 | 0.7450 | 0.8990 |
| | Outubro Rosa | 0.9300 | 0.9320 | 0.9120 | 0.4540 | 0.7100 |
| Canada | Novembro Azul | 0.9960 | 0.9820 | 0.9920 | 0.9890 | 0.9940 |
| | Outubro Rosa | 0.9530 | 0.9750 | 0.9850 | 0.9690 | 0.9800 |
| Mexico | Novembro Azul | 0.9510 | 0.9640 | 0.9450 | 0.9470 | 0.9850 |
| | Outubro Rosa | 0.9680 | 0.9850 | 0.9040 | 0.9500 | 0.9430 |
| United Kingdom | Novembro Azul | 0.9970 | 0.9940 | 0.9950 | 0.9960 | 0.9990 |
| | Outubro Rosa | 0.9360 | 0.9680 | 0.9760 | 0.9620 | 0.9640 |
| United States | Novembro Azul | 0.9900 | 0.9880 | 0.9660 | 0.9840 | 0.9890 |
| | Outubro Rosa | 0.9740 | 0.9960 | 0.9950 | 0.9900 | 0.9840 |

Figura 5.28: Índices de correlação entre as distribuições de postagem por data - geral vs. detalhada por pais

Concluímos que as as campanhas possuem ações ao longo do mês similares entre os países. Os países apresentam o mesmo padrão de comportamento nas respectivas campanhas em cada ano. No entanto, o Brasil aparece como exceção neste cenário, apresentando algumas evoluções distintas dos demais países, em especial no ano de 2017 na campanha do Outubro Rosa.

### 5.5.3 Abrangência

Na Seção 5.3 concluímos que as campanhas do OR são bem mais ativas que as do NA, tanto em número de postagens quanto de propagação (*retweets*). Como para os anos de 2014-2016 não possuímos informações detalhadas sobre o comportamento *retweets* devido à ferramenta de coleta, e dada a diferença no volume de *tweets* nos diferentes anos, nossa análise longitudinal foi limitada às médias de postagem e republicações.

A Figura 5.29 resume para cada ano e campanha, a média de *tweets* por categoria de usuário. As cores laranja são utilizada para destacar as menores médias, e as azuis para salientar a maiores média. Quanto maior/menor valor, mais intensa é a cor utilizada. Por exemplo, as organizações possuem a maior média, e de uma forma geral, as celebridades as menores médias. O valor mais baixo é de 1,39 (média de *tweets* para celebridades no Novembro Azul, no ano de 2015), e o valor mais alto é de 3,64 (média de *tweets* para organizações no Outubro Rosa, também no ano de 2015).

Com base na Figura 5.29, com algumas exceções confirmamos de uma forma geral os padrões encontrados para o ano de 2017 nos demais anos: a) as organizações são as que possuem maiores médias de *tweets*, b) os indivíduos tendem a postar mais que as celebridades. Isso mostra que as celebridades não desempenham o papel proeminente que se espera delas, principalmente na campanha do Outubro Rosa. As celebridades estão mais envolvidas no Novembro Azul, quando foram em média mais ativas que os indivíduos nos anos de 2016 e 2018, ou atuaram em níveis relativamente próximos (2014 e 2017). Já as organizações estão mais envolvidas no Outubro Rosa, com médias variando entre 2,98 e 3,64 *tweets*, enquanto que no Novembro Azul a média máxima observada é de 2,77 no ano de 2014. Esses pequenos fatores, na escala do tamanho dessas campanhas no Twitter, podem fazer bastante diferença nos resultados de engajamento nestas campanhas.

| | | Ano | | | | |
|---|---|---|---|---|---|---|
| | | 2014 | 2015 | 2016 | 2017 | 2018 |
| Novembro Azul | Celebridade | 1,83 | 1,39 | 2,06 | 1,58 | 2,23 |
| | Indivíduo | 1,83 | 1,87 | 1,81 | 1,64 | 1,98 |
| | Organização | 2,77 | 2,51 | 2,06 | 2,33 | 2,65 |
| Outubro Rosa | Celebridade | 1,70 | 1,67 | 1,56 | 1,53 | 1,79 |
| | Indivíduo | 1,85 | 1,95 | 2,02 | 2,17 | 1,78 |
| | Organização | 2,98 | 3,64 | 3,12 | 3,53 | 3,03 |

Figura 5.29: Relação de número de usuários, *tweets* e média de *tweets* por tipo de usuário

Na Figura 5.30 fazemos uma avaliação da abrangência que os *tweets* atingem na rede por meio de *retweets* representando a distribuição de *retweets* utilizando *boxplots*. Identificamos anteriormente (Figuras 5.12 e 5.13), que a campanha do Outubro Rosa apresenta resultados um pouco melhores que o Novembro Azul, com os *tweets* desta campanha sendo mais propagados na rede social. Avaliando ao longo dos anos, podemos entender que ambas as campanhas apresentam uma evolução gradativa ao longo dos anos, tanto quer em termos de medianas quanto do valor do terceiro quartil. O ano de 2018 teve uma alta acima do padrão de crescimento, já que no Outubro Rosa as medianas de *retweets* mais que dobraram, além de elevação significativa do terceiro quartil.

Ao avaliarmos esse comportamento por país na Figura 5.31, podemos perceber que ele ocorreu em maior ou menor escala em todos os países. O destaque no crescimento de *retweets* é o Brasil. Podemos perceber que este comportamento ocorreu também no que tange o Novembro Azul, mas em uma escala menor que do Outubro Rosa.

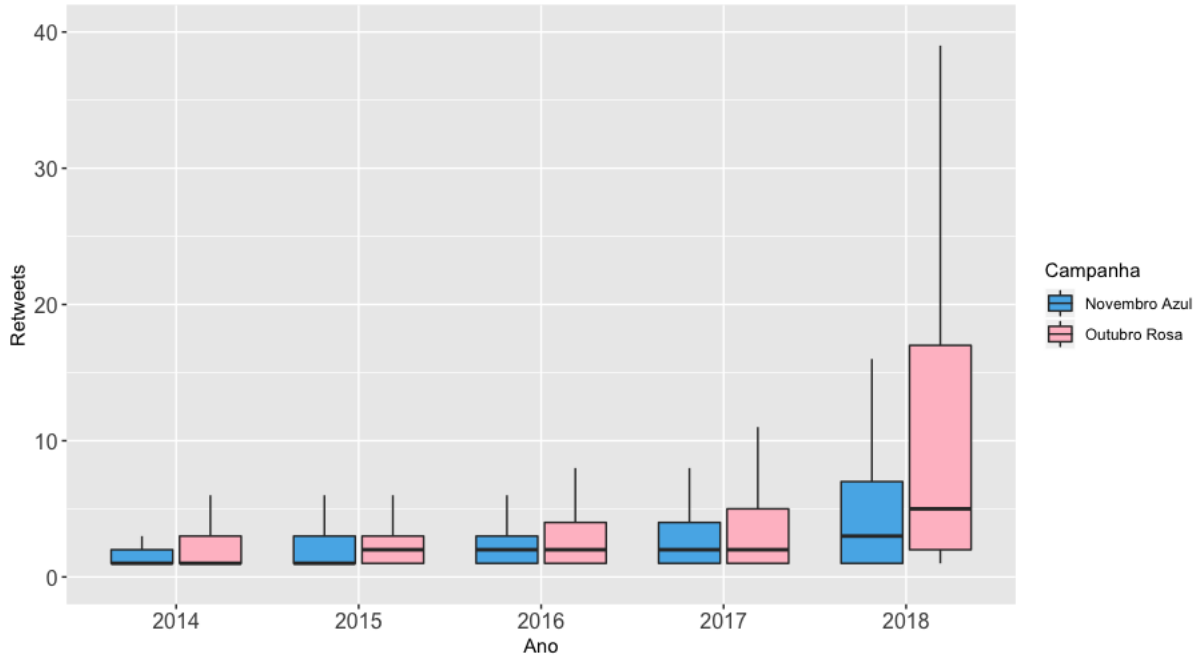


Figura 5.30: Distribuição da frequência de *retweets* por campanha e ano

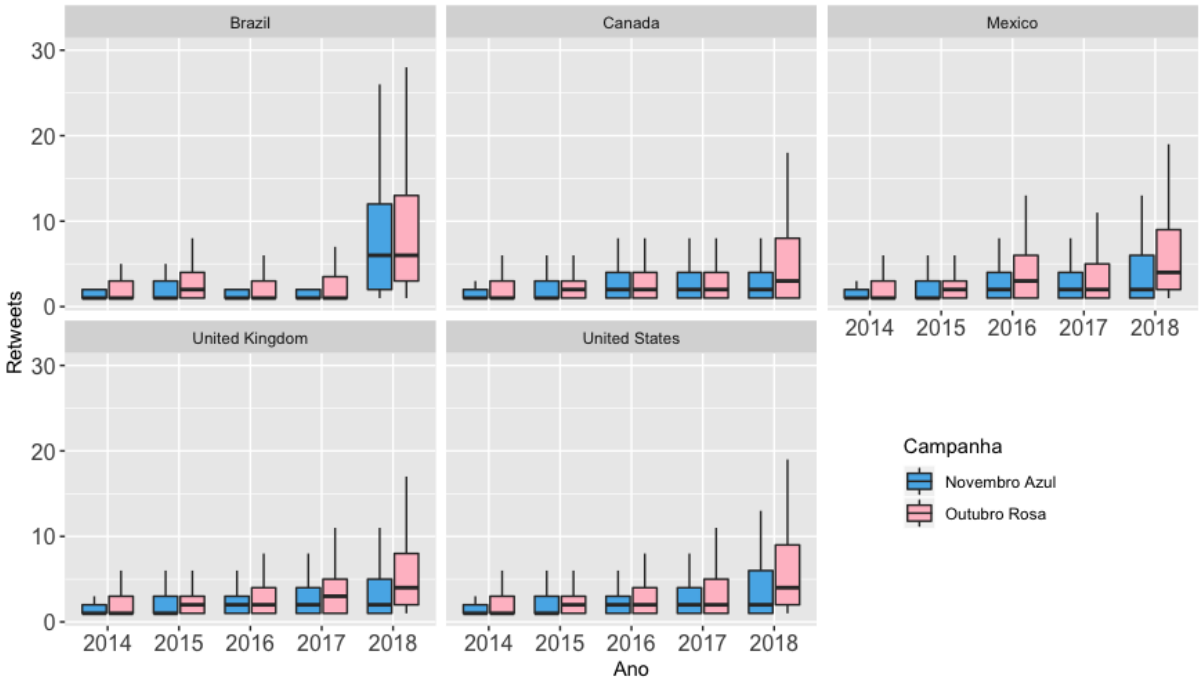


Figura 5.31: Distribuição da frequência de *retweets* por campanha, ano e país

Podemos concluir que as campanhas vem apresentando uma evolução gradativa de abrangência e propagação, e que o Novembro Azul não possui o mesmo alcance do Outubro Rosa. O Brasil tem apresentando uma evolução maior que os demais países, especialmente no ano de 2018. Entendemos também que as celebridades não desempenham o papel que se espera delas, e uma vez que as mesmas estejam mais engajadas, as duas campanhas aumentam sua capacidade de propagação de informações.

### 5.5.4 Tópicos

Finalmente, comparamos os tópicos discutidos em cada campanha ao longo dos diferentes anos. A Figura 5.32 mostra a representatividade de cada tópico em cada campanha ao longo dos anos de 2014 a 2018.

Com base na Figura 5.32, percebemos diferenças significativas entre a representatividade das diferentes categorias. Identificamos que a categoria *Fundraising* é que a apresenta maiores índices de participação em todos os anos. Ela é seguida pela categoria *Wear Pink/Blue*, que também é bem significativa, mas cuja representatividade vem decrescendo ao longo dos anos, principalmente para o Outubro Rosa. As demais categorias apresentam oscilações em sua representatividade.

Quanto à categoria *Loved Ones*, exceto pelo ano de 2015, sempre está presente no Outubro Rosa, com valores oscilando entre 10% e 30%. Contudo, ela só foi observado no Novembro Azul a partir de 2018. Oscilações similares ocorrem com a categoria *Early Detection/Diagnosis*, onde apenas em 2017 ela tem representação na campanha do Novembro Azul, e nos demais anos a categoria participa somente no Outubro Rosa.

Nota-se uma representatividade crescente de tópicos classificados como *Others* no Outubro Rosa, sugerindo uma evolução nos tópicos ou nos termos utilizados. Este fenômeno é contrário no caso do Novembro Azul, sendo observado apenas nos anos de 2014 e 2015.

A Figura 5.33 detalha esta informação por país. Observa-se que os EUA, Canadá e Reino Unido mantêm um comportamento muito semelhante, no sentido de que o aumento/diminuição na prevalência de um tópico é comum a todos eles. Por exemplo, a prevalência da categoria *Fundraising* na campanha Outubro Rosa, foi diminuindo gradativamente em todos os países até o ano de 2016, tendo aumentado em 2017, e com exceção dos EUA, voltou a diminuir em 2018. Para a campanha Novembro Azul, com exceção do ano 2016 no Reino Unido, o comportamento desta mesma categoria em todos os anos de crescimento ou diminuição do tópico foi muito similar nos diferentes países.

Outro padrão interessante é que quando um tópico foi detectado, ele esteve presente em todos os países. Este é o caso da categoria *Treatments*, que foi identificada somente nos anos de 2016 e 2017 na campanha do Outubro Rosa em todos os países. Em 2018 não foram identificados tópicos relacionados a esta categoria. Isso pode indicar que os 3 países estão alinhados nos conteúdos das campanhas, e talvez são influenciados uns pelos outros.

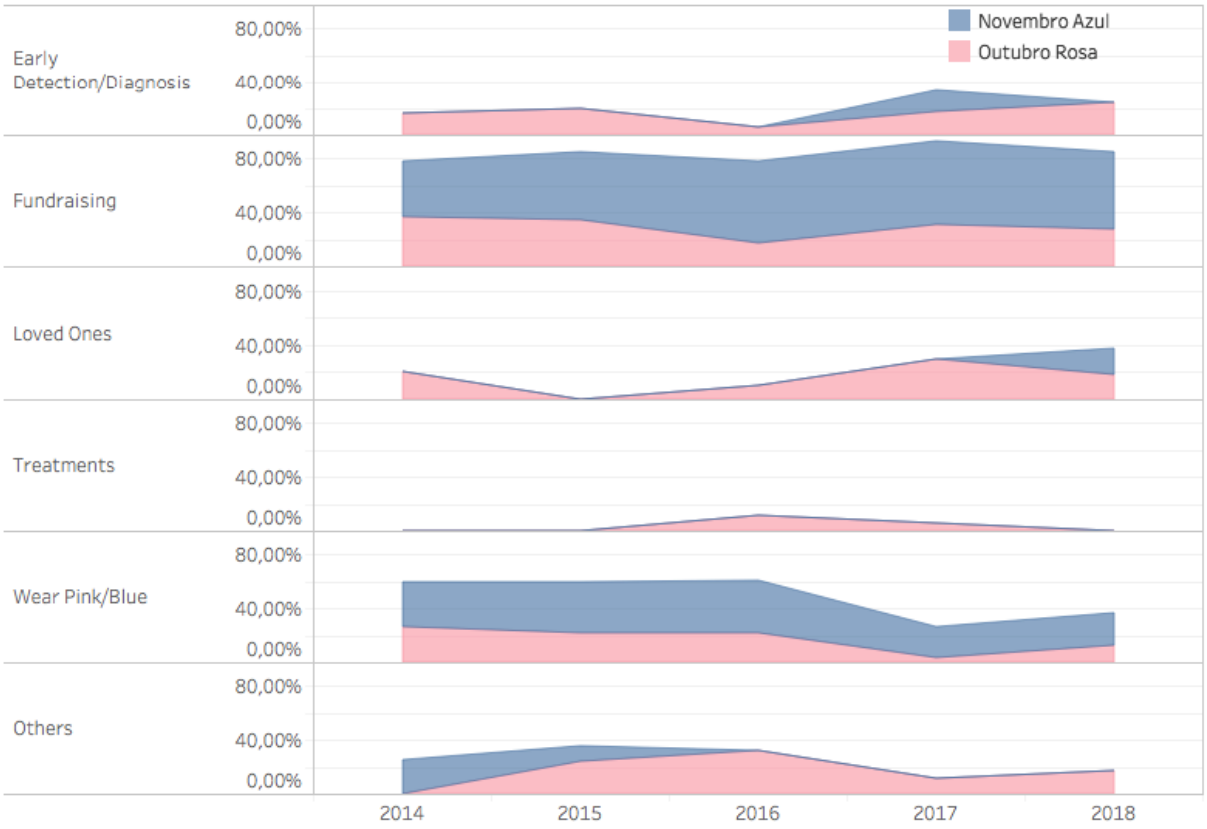


Figura 5.32: Evolução das categorias dos tópicos ao longo dos anos

Concluímos que as campanhas possuem diferentes perspectivas nos assuntos abordados. O Novembro Azul, talvez por estar em um estado diferente de maturidade do Outubro Rosa, está principalmente focado em levantar recursos financeiros e divulgar a campanha. O Outubro Rosa, por outro lado, além de ter a preocupação financeira e de divulgação da campanha, também emprega sua atenção nas atividades de detecção do câncer nos estágios iniciais, e nos cuidados com os entes queridos. Identificamos que estes padrões tem se mantido ao longo dos anos.

## 5.6 Considerações Finais

Neste capítulo foram apresentadas as análises relativas às questões de pesquisa propostas. Inicialmente, foi verificado que no tocante ao público, as campanhas do Outubro Rosa e Novembro Azul no Twitter atingem o público alvo. Identificamos que as campanhas podem ter variações de comportamento em diferentes países, bem como diferenças de comportamento entre o Outubro Rosa e Novembro Azul dentro de alguns países. No tocante à abrangência, concluímos que o Novembro Azul não possui o mesmo alcance do Outubro Rosa, e que as celebridades não desempenham um papel proeminente que poderiam ter nestas campanhas. Por fim,

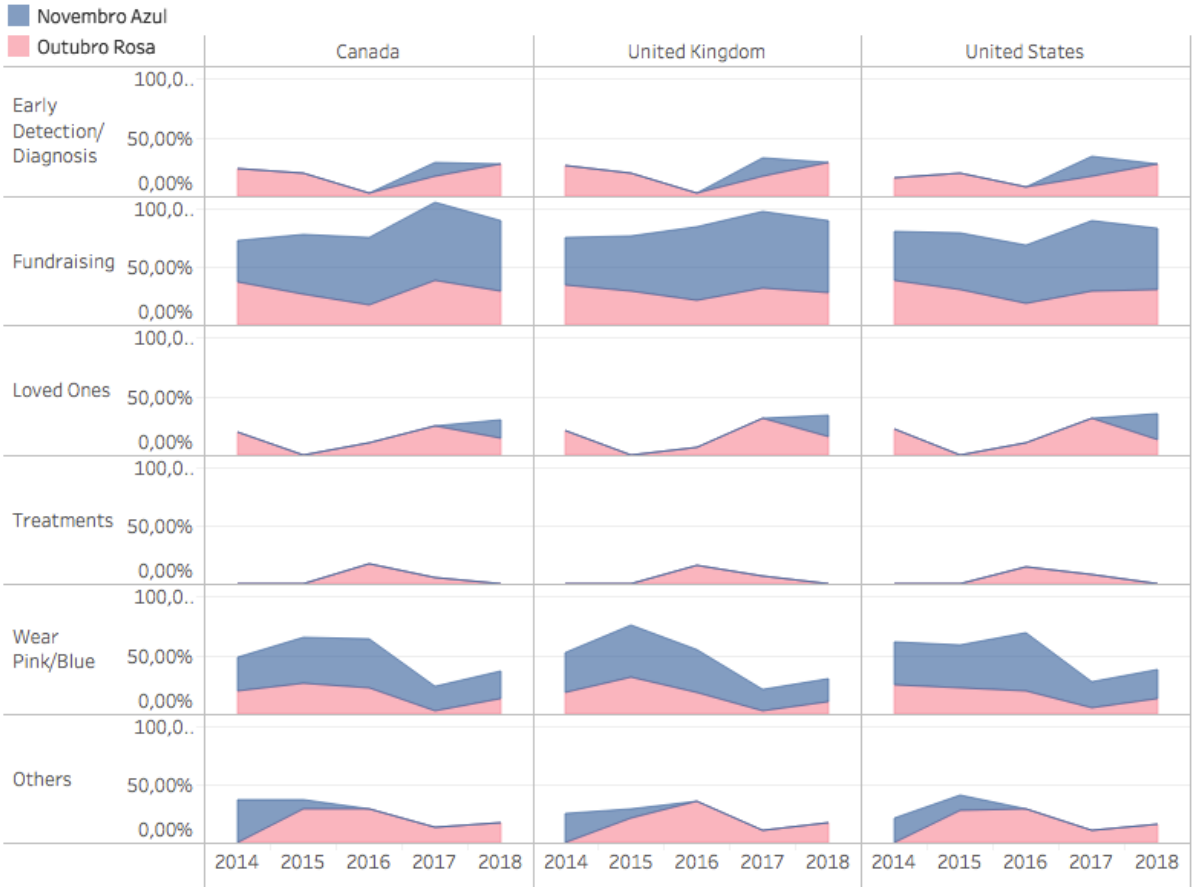


Figura 5.33: Evolução das categorias nos países ao longo dos anos

identificamos que os *tweets* do Novembro Azul estão mais focados em levantar recursos e em propagar a campanha em si, enquanto que o Outubro Rosa abrange uma variedade de assuntos, desde a propagação de informações, até a preocupação com os entes queridos e realização de exames de detecção do câncer. A avaliação longitudinal revelou que estes padrões se mantém de uma forma geral, mas que existe uma disparidade no volume de participação que pode ter influenciado os demais indicadores.

Concluímos o capítulo discutindo ameaças que influenciam a validade de nossas análises. A inferência das informações demográficas é uma das principais ameaças para nossas análises de dados demográficos. Experimentos com o Face++ apontam acurácia mínima de 85% Fan et al. (2014), mas os usuários podem adotar imagens de perfil que não representam sua aparência. Neste trabalho, o Face++ não inferiu as informações demográficas sobre 53% do nosso conjunto de usuários. Uma segunda ameaça é a definição de país do usuário para as nossas análises de dados geográficos. Dado que o número de *tweets* geo-localizados é muito pequeno (menos de 2%), utilizamos a informação de localidade informada no perfil do usuário. No entanto, essa informação em muitos casos é imprecisa e não representa uma localização válida. Utilizamos a ferramenta do Google Maps para validar essa informação e produzir resul-

tados mais confiáveis.

Por fim, reconhecemos que o conjunto de *hashtags* pode apresentar limitações. Neste caso, o Novembro Azul contém um conjunto menor de *hashtags* comparado ao Outubro Rosa, e que podem comprometer na representação de dados objetos das nossas análises, como por exemplo, a falta de representação de *tweets* sobre uma categoria de tópico.

# 6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho propôs uma avaliação comparativa do comportamento no Twitter em relação às campanhas do Outubro Rosa e Novembro Azul no Twitter. Comparamos as características demográficas dos usuários participantes, as atividade ao longo da campanha e os principais assuntos abordados. A análise partiu de um ano específico (2017), e avaliou as diferenças dos padrões encontrados em relação a anos anteriores (2014-2017) e ano seguinte (2018). Para viabilizar a análise e interpretação dos tópicos, considerando duas campanhas, 3 países e 5 anos, propusemos um método baseado na adaptação de um referencial de categorias de assuntos existentes, e na proximidade de termos representados através de *word embeddings* contidos nos *tweets* e neste referencial. Foram apresentados experimentos com o objetivo de comparar o presente trabalho ao estado-da-arte e avaliar os componentes e estratégias utilizadas neste processo.

Como resultado deste trabalho, providenciamos algumas intuições sobre uma comparação das campanhas do Outubro Rosa e Novembro Azul no Twitter, resultando em boas reflexões de múltiplos aspectos das campanhas. Nossas análises mostram que o público alvo das campanhas (homens e mulheres acima dos 40 anos) é o que mais engaja, e que existem diferenças nessa participação nos diferentes países, principalmente no Brasil. Identificamos também que as estruturas de rede e engajamento da campanha do Outubro Rosa são maiores que a do Novembro Azul, o que resulta da maior propagação da primeira campanha na rede. Por último, também identificamos que a campanha do Outubro Rosa aparenta estar mais madura nos assuntos que são abordados nos *tweets*, pois os participantes abordam desde o levantamento de recursos até relatos sobre os seus entes queridos. O Novembro Azul talvez esteja em um momento de preocupação com a expansão da campanha, o que pode estar direcionando esta campanha a focar principalmente em aspectos financeiros e de propagação da mesma. Esperamos que este trabalho contribua com reflexões que levem ao progresso nas campanhas do Novembro Azul, e que as mesmas possam alcançar números expressivos como as campanhas do Outubro Rosa.

Em nosso entendimento, trata-se de um trabalho pioneiro, já que os trabalhos existentes limitam a análise a uma única campanha (THACKERAY et al., 2013; NASTASI et al., 2017; PRASETYO et al., 2015; JACOBSON; MASCARO, 2016). Ainda, detalha vários aspectos anteriormente não abordados por completo, tais como tipo de usuário, dados geográficos e demográficos dos usuários, análise de conteúdo e comparação entre campanhas, que já haviam sido abordados parcialmente em outros domínios (BORGMANN et al., 2016; GLYNN et al., 2011; THACKERAY et al., 2013; NASTASI et al., 2017; PRASETYO et al., 2015).

Em relação à identificação de tópicos, contribuímos ao posicionar as categorias identificadas por Thackeray et al. (2013) em um contexto mais amplo, e utilizando *word embeddings* como forma de viabilizar a interpretação de tópicos em um estudo longitudinal envolvendo três países. Neste sentido, nosso método de interpretação e categorização dos tópicos necessita de uma interferência mínima, apenas para definir o conjunto inicial de palavras e categorias que compõem o referencial de interpretação. Após essa definição, o método não requer mais interferências manuais. Assim, o método pode ser facilmente replicado para outras campanhas além das que foram objetos deste estudo.

O trabalho de pesquisa desenvolvido resultou até o momento em duas publicações:

- Walter, R. e Becker, K. (2018). Caracterização e Comparação das Campanhas do Outubro Rosa e Novembro Azul no Twitter. Em **Anais do Simpósio Brasileiro de Banco de Dados 2018**, Rio de Janeiro, RJ., 2018.

- Walter, R. e Becker, K. (2018). Caracterização e Comparação de Campanhas Promovendo o Outubro Rosa e o Novembro Azul no Twitter. Em **Anais do SBBD/WTDBD - XV Workshop de Teses e Dissertações em Banco de Dados**. Rio de Janeiro, RJ., 2018.

Identificamos no trabalho algumas limitações, e que nos dão uma perspectiva de continuação para evolução deste trabalho:

- A recuperação do gênero e idade com a ferramenta Face++ possui baixos índices de recuperação das informações, onde aproximadamente $47\%$ dos usuários foram classificados de acordo com os critérios de gênero/idade;

- As palavras dos tópicos podem mudar ao longo dos anos, então uma categoria pode ficar desatualizada;

- Falta de referência de *retweets* nos dados coletados com a ferramenta GetOldTweets-python[1];

- A avaliação das categorias dos *tweets* foi feita somente no idioma inglês. Para contemplar todos os *tweets* coletados, é necessário expandir essa aplicação para *tweets* em português e espanhol;

- A categoria *Others* não foi detalhada. É possível que a partir de uma avaliação minuciosa da mesma surjam novas categorias de tópicos.

[1] https://github.com/Jefferson-Henrique/GetOldTweets-python

O trabalho proposto pode evoluir a fim de remover estas limitações. Para isso, trabalhos futuros podem explorar: $a$) novos formatos de recuperação de informações demográficas, a fim de aumentar os índices de recuperação dessas informações, $b$) avaliação de mudança de vocabulário e *topic drifting* para avaliar a evolução das categorias dos tópicos, $c$) unificar a coleta dos *tweets* dos próximos anos na integração nativa do Twitter, que retorna as referências de *retweets*, $d$) evoluir a categorização e geração de tópicos nos idiomas português e espanhol, e $e$) detalhar a categoria de tópicos *Others* a fim de identificar potenciais novas categorias de tópicos para os *tweets*. Estes itens geram perspectivas para melhorar o trabalho, mas é importante ressaltar que as análises e o desenvolvimento do método de categorização dos *tweets* foram realizados conforme pretendido.

# REFERÊNCIAS

AGGARWAL, C. C.; ZHAI, C. A survey of text clustering algorithms. In: **Mining Text Data**. [s.n.], 2012. p. 77–128. Available from Internet: <https://doi.org/10.1007/978-1-4614-3223-4_4>.

ALGHAMDI, R.; ALFALQI, K. A survey of topic modeling in text mining. **Int. J. Adv. Comput. Sci. Appl.(IJACSA)**, Citeseer, v. 6, n. 1, 2015.

ALSUMAIT, L. et al. Topic significance ranking of lda generative models. In: SPRINGER. **Joint European Conference on Machine Learning and Knowledge Discovery in Databases**. [S.l.], 2009. p. 67–82.

ALTEKRUSE, S. et al. Seer cancer statistics review, 1975–2007. **Bethesda, MD: National Cancer Institute**, v. 7, 2010.

BAEZA-YATES, R.; RIBEIRO-NETO, B. et al. **Modern information retrieval**. [S.l.]: ACM press New York, 1999.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003.

BORGMANN, H. et al. Activity, content, contributors, and influencers of the twitter discussion on urologic oncology. In: ELSEVIER. **Urologic Oncology: Seminars and Original Investigations**. [S.l.], 2016. v. 34, p. 377–383.

BRAVO, C. A.; HOFFMAN-GOETZ, L. Tweeting about prostate and testicular cancers: Do twitter conversations and the 2013 movember canada campaign objectives align? **Journal of Cancer Education**, Springer, v. 31, n. 2, p. 236–243, 2016.

CASTLETON, K. et al. A survey of internet utilization among patients with cancer. **Supportive Care in Cancer**, Springer, v. 19, n. 8, p. 1183–1190, 2011.

CHANG, J. et al. Reading tea leaves: How humans interpret topic models. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2009. p. 288–296.

CHOUDHURY, M. D. et al. Social media participation in an activist movement for racial equality. In: **ICWSM**. [S.l.: s.n.], 2016. p. 92–101.

CONTROL, C. for D.; PREVENTION. 2012. 61(03); 41–45 p. Available from Internet: <https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6103a1.htm>.

DEERWESTER, S. et al. Indexing by latent semantic analysis. **Journal of the American society for information science**, Wiley Online Library, v. 41, n. 6, p. 391–407, 1990.

DENNY, M. J.; SPIRLING, A. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. **Political Analysis**, Cambridge University Press, v. 26, n. 2, p. 168–189, 2018.

DOUVEN, I.; MEIJS, W. Measuring coherence. **Synthese**, Springer, v. 156, n. 3, p. 405–425, 2007.

EDGE, L. Breast-cancer awareness: too much of a good thing? **Lancet Oncol**, v. 7, p. 611, 2006.

ELSHERIEF, M.; BELDING, E. M.; NGUYEN, D. # notokay: Understanding gender-based violence in social media. In: **ICWSM**. [S.l.: s.n.], 2017. p. 52–61.

ESTÉVEZ, P. A. et al. Normalized mutual information feature selection. **IEEE Transactions on Neural Networks**, IEEE, v. 20, n. 2, p. 189–201, 2009.

FAN, H. et al. Learning deep face representation. **arXiv preprint arXiv:1403.2802**, 2014.

FOUNDATION, I. N. B. C. 2017. Available from Internet: <http://www.nationalbreastcancer.org/breast-cancer-awareness-month>.

GLYNN, R. W. et al. The effect of breast cancer awareness month on internet search activity-a comparison with awareness campaigns for lung and prostate cancer. **BMC cancer**, BioMed Central, v. 11, n. 1, p. 442, 2011.

GOLDBERG, Y.; LEVY, O. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. **arXiv preprint arXiv:1402.3722**, 2014.

HARB, J. G. D.; BECKER, K. Emotion analysis of reaction to terrorism on twitter. In: **Proc. of the SBC Brazilian Symposium on Databases**. [S.l.: s.n.], 2018. p. 97–108.

HIMELBOIM, I.; HAN, J. Y. Cancer talk on twitter: community structure and information sources in breast and prostate cancer social networks. **Journal of health communication**, Taylor & Francis, v. 19, n. 2, p. 210–225, 2014.

HOFMANN, T. Probabilistic latent semantic analysis. In: MORGAN KAUFMANN PUBLISHERS INC. **Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence**. [S.l.], 1999. p. 289–296.

JACOBSEN, G. D.; JACOBSEN, K. H. Health awareness campaigns and diagnosis rates: evidence from national breast cancer awareness month. **Journal of health economics**, Elsevier, v. 30, n. 1, p. 55–61, 2011.

JACOBSON, J.; MASCARO, C. Movember: Twitter conversations of a hairy social movement. **Social Media+ Society**, SAGE Publications Sage UK: London, England, v. 2, n. 2, p. 2056305116637103, 2016.

LAFFERTY, J. D.; BLEI, D. M. Correlated topic models. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2006. p. 147–154.

LARANJO, L. et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 22, n. 1, p. 243–256, 2014.

LEVY, O.; GOLDBERG, Y. Dependency-based word embeddings. In: **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. [S.l.: s.n.], 2014. v. 2, p. 302–308.

LI, Q. et al. Real-time novel event detection from social media. In: **Proc. of the 33rd IEEE International Conference on Data Engineering (ICDE)**. [S.l.: s.n.], 2017. p. 1129–1139.

LOTAN, G. et al. The arab spring| the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. **International journal of communication**, v. 5, p. 31, 2011.

MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to information retrieval. **Natural Language Engineering**, Cambridge university press, v. 16, n. 1, p. 100–103, 2010.

MCCARTNEY, M. Is movember misleading men? **Bmj**, British Medical Journal Publishing Group, v. 345, p. e8046, 2012.

MEGVII, I. **Face++ research toolkit**. 2013.

MEHROTRA, R. et al. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In: ACM. **Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval**. [S.l.], 2013. p. 889–892.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2013. p. 3111–3119.

MIMNO, D.; BLEI, D. Bayesian checking for topic models. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the conference on empirical methods in natural language processing**. [S.l.], 2011. p. 227–237.

MISLOVE, A. et al. Understanding the demographics of twitter users. In: **Fifth international AAAI conference on weblogs and social media**. [S.l.: s.n.], 2011.

NASTASI, A. et al. Breast cancer screening and social media: a content analysis of evidence use and guideline opinions on twitter. **Journal of Cancer Education**, Springer, p. 1–8, 2017.

NETO, J. A. M.; BECKER, K. Relating conversational topics and toxic behavior effects in a MOBA game. **Entertainment Computing**, v. 26, p. 10–29, 2018. Available from Internet: <https://doi.org/10.1016/j.entcom.2017.12.004>.

OLTEANU, A.; WEBER, I.; GATICA-PEREZ, D. Characterizing the demographics behind the# blacklivesmatter movement. **OSSM. http://arxiv. org/abs/1512.05671**, 2016.

PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: Global vectors for word representation. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543.

PRASETYO, N. D. et al. On the impact of twitter-based health campaigns: A cross-country analysis of movember. In: **Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis**. [S.l.: s.n.], 2015. p. 55–63.

QI, Y. et al. When and why are pre-trained word embeddings useful for neural machine translation? **arXiv preprint arXiv:1804.06323**, 2018.

RAJARAMAN, A.; ULLMAN, J. D. **Mining of massive datasets**. [S.l.]: Cambridge University Press, 2011.

RESNIK, P. et al. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In: **Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality**. [S.l.: s.n.], 2015. p. 99–107.

RESNIK, P.; GARRON, A.; RESNIK, R. Using topic modeling to improve prediction of neuroticism and depression in college students. In: **Proceedings of the 2013 conference on empirical methods in natural language processing**. [S.l.: s.n.], 2013. p. 1348–1353.

RÖDER, M.; BOTH, A.; HINNEBURG, A. Exploring the space of topic coherence measures. In: ACM. **Proceedings of the eighth ACM international conference on Web search and data mining**. [S.l.], 2015. p. 399–408.

SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake shakes twitter users: Real-time event detection by social sensors. In: **Proc. of the 19th Intl. Conf. on World Wide Web (WWW)**. [S.l.: s.n.], 2010. p. 851–860.

SCHNABEL, T. et al. Evaluation methods for unsupervised word embeddings. In: **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2015. p. 298–307.

SCHULZ, A.; SCHMIDT, B.; STRUFE, T. Small-scale incident detection based on microposts. In: ACM. **Proceedings of the 26th ACM Conference on Hypertext & Social Media**. [S.l.], 2015. p. 3–12.

SLOAN, L. et al. Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. **PloS one**, Public Library of Science, v. 10, n. 3, p. e0115545, 2015.

SURIAN, D. et al. Characterizing twitter discussions about hpv vaccines using topic modeling and community detection. **Journal of medical Internet research**, JMIR Publications Inc., Toronto, Canada, v. 18, n. 8, p. e232, 2016.

THACKERAY, R. et al. Using twitter for breast cancer prevention: an analysis of breast cancer awareness month. **BMC cancer**, BioMed Central, v. 13, n. 1, p. 508, 2013.

WALTER, R.; BECKER, K. Caracterização e comparação das campanhas do outubro rosa e novembro azul no twitter. In: **Proc. of the SBC Brazilian Symposium on Databases**. [S.l.: s.n.], 2018. p. 133–144.

ZHAO, W. X. et al. Comparing twitter and traditional media using topic models. In: SPRINGER. **European conference on information retrieval**. [S.l.], 2011. p. 338–349.

# Appendices

# ApêndiceA

## A.1 Matrizes de similaridade das relações entre categorias e tópicos do Novembro Azul

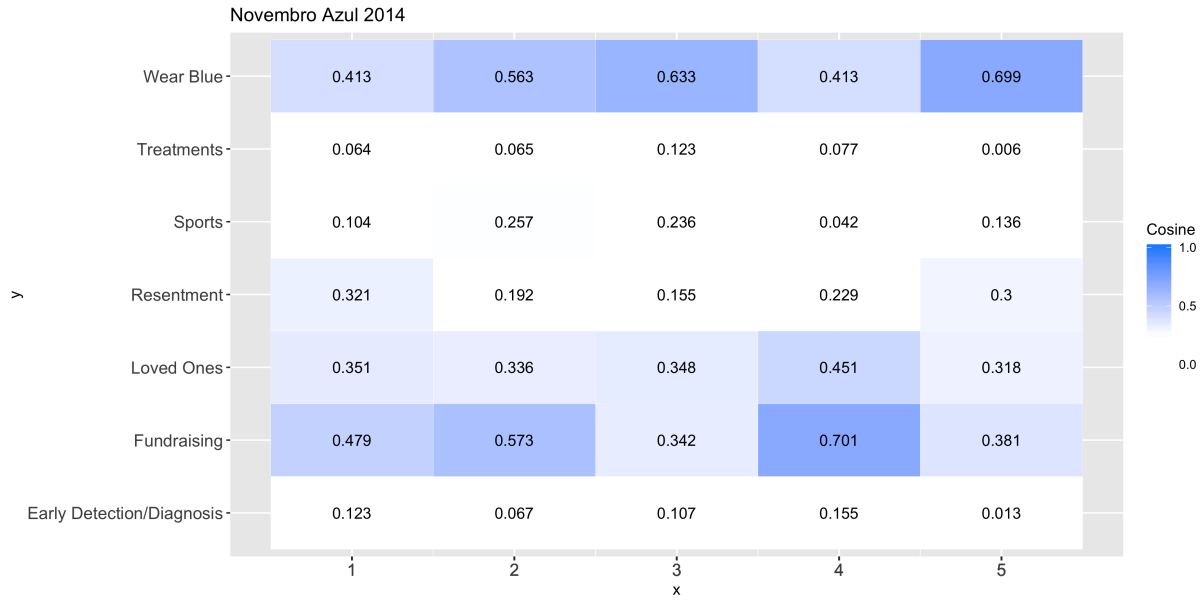


Figura A.1: Matriz de similaridade da relação entre categorias e tópicos do Novembro Azul do ano de 2014

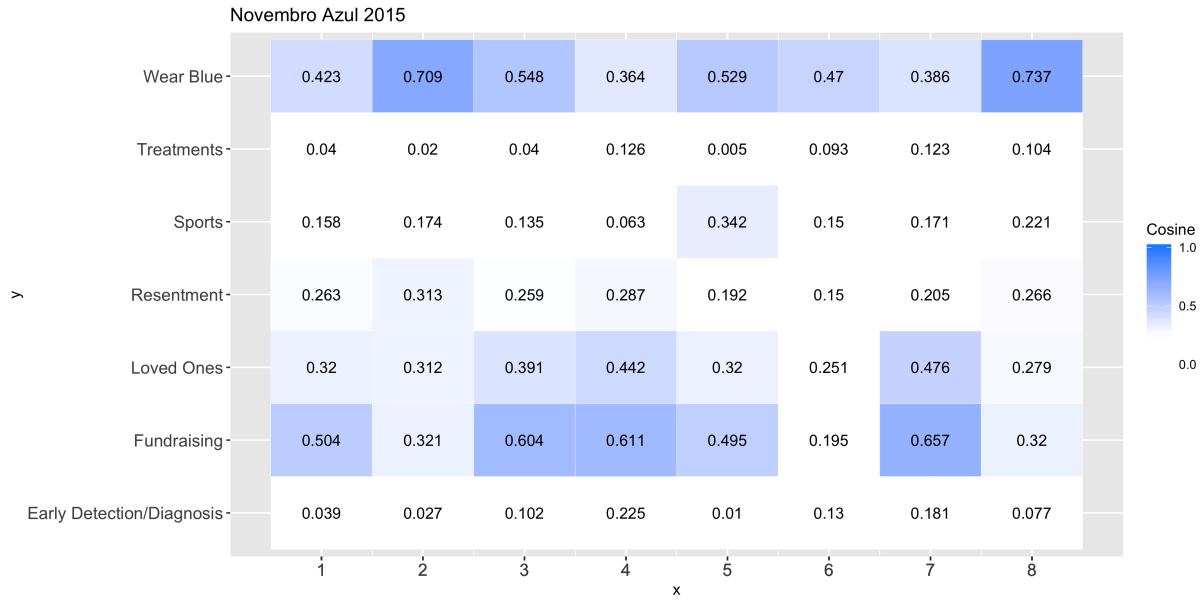


Figura A.2: Matriz de similaridade da relação entre categorias e tópicos do Novembro Azul do ano de 2015

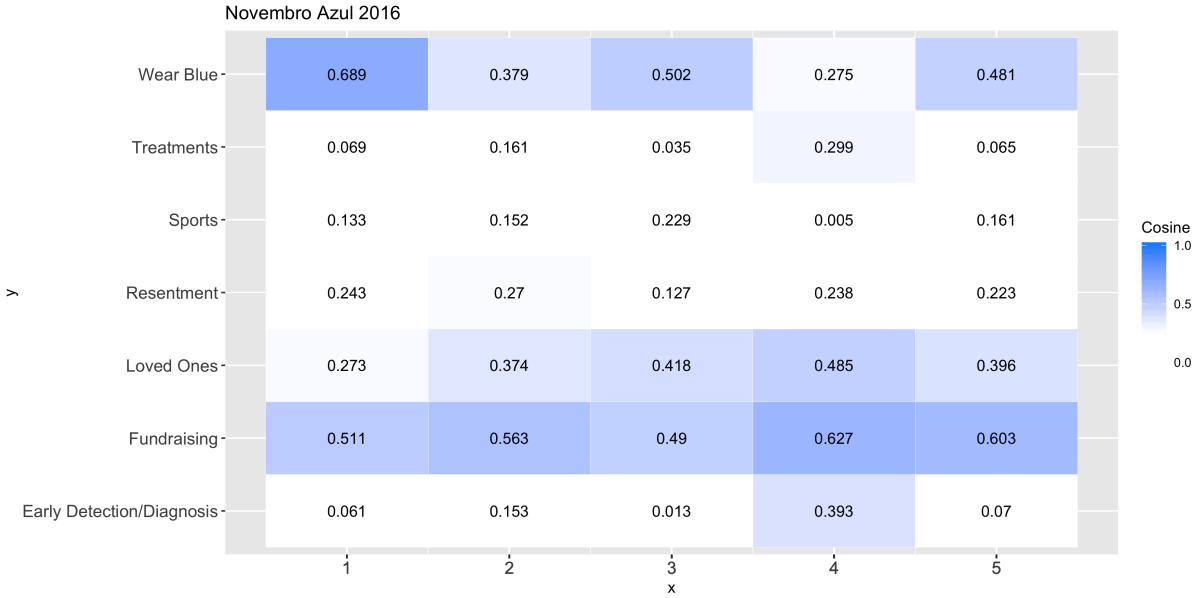


Figura A.3: Matriz de similaridade da relação entre categorias e tópicos do Novembro Azul do ano de 2016

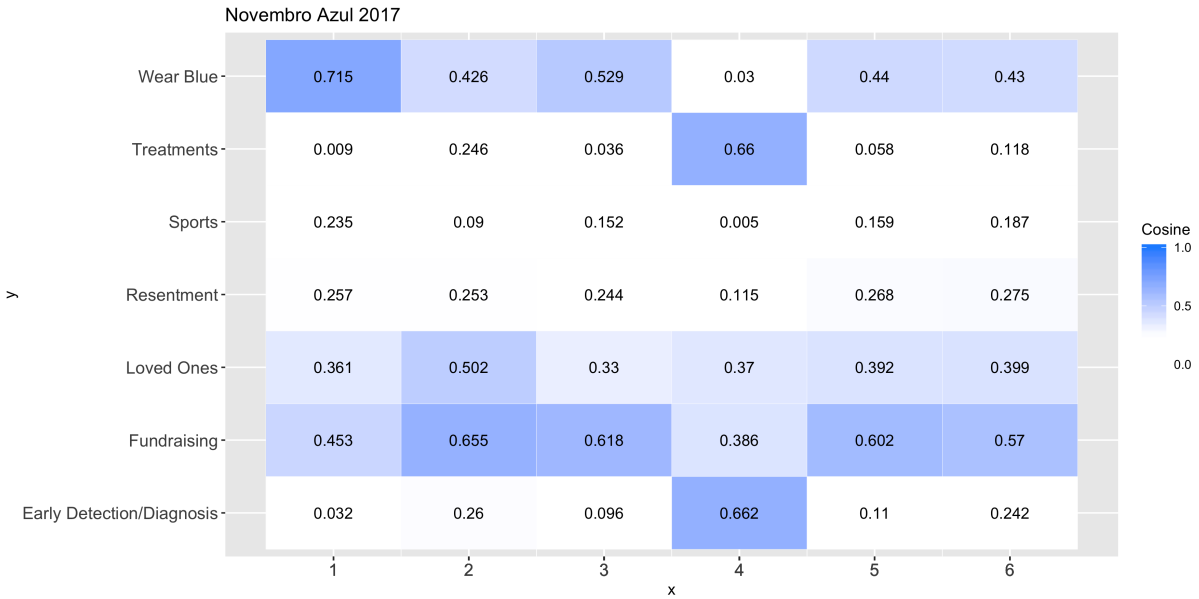


Figura A.4: Matriz de similaridade da relação entre categorias e tópicos do Novembro Azul do ano de 2017

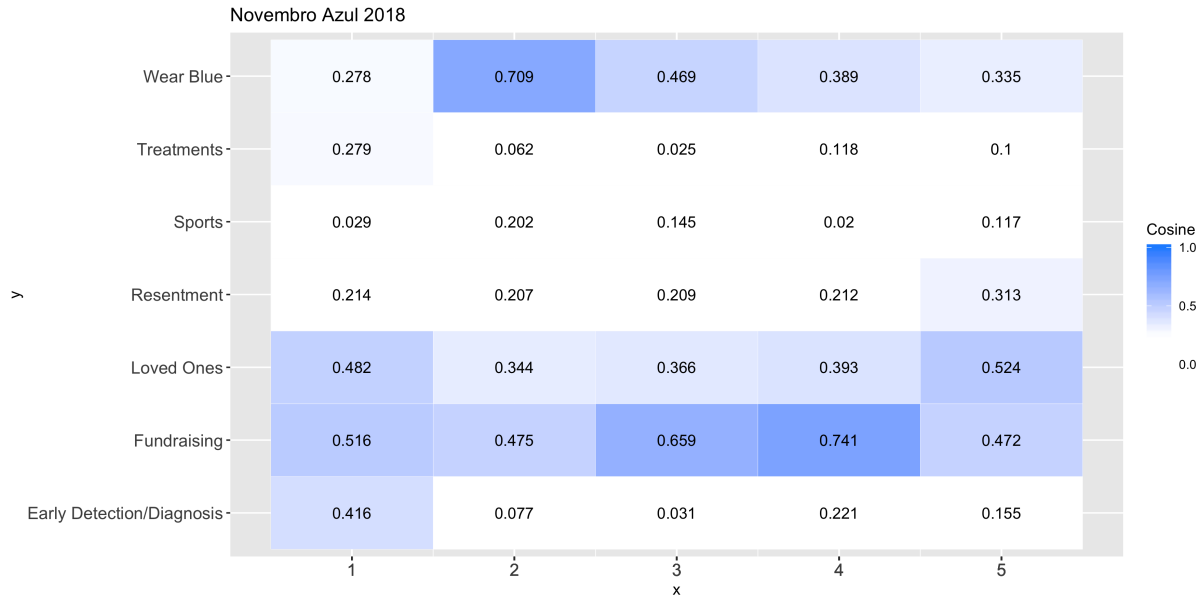


Figura A.5: Matriz de similaridade da relação entre categorias e tópicos do Novembro Azul do ano de 2018

## A.2 Matrizes de similaridade das relações entre categorias e tópicos do Outubro Rosa

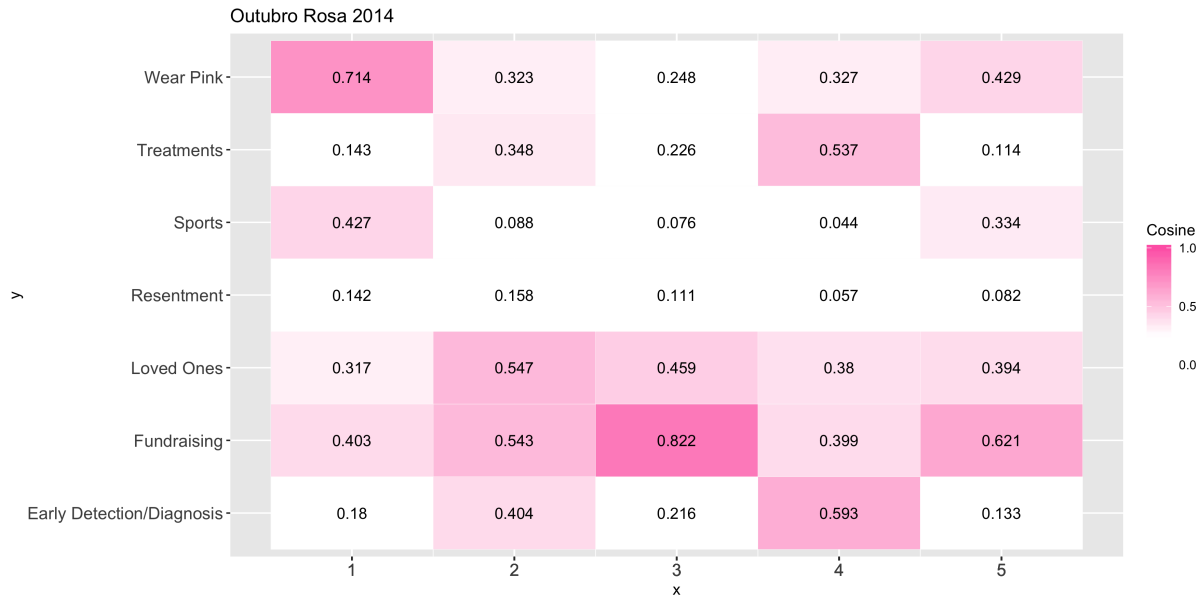


Figura A.6: Matriz de similaridade da relação entre categorias e tópicos do Outubro Rosa do ano de 2014

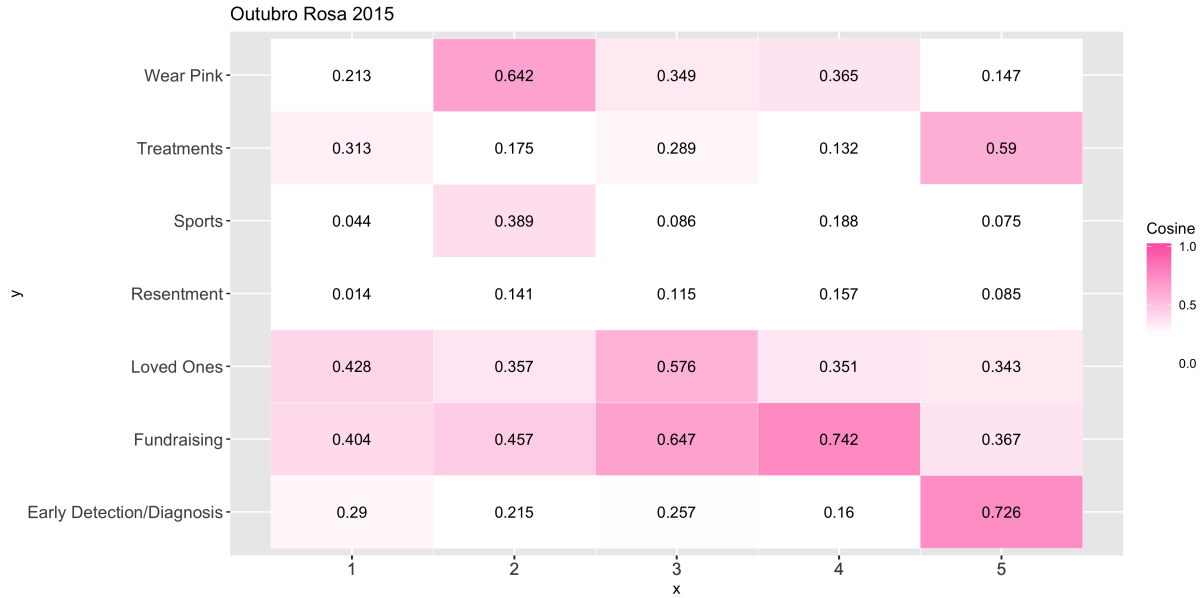


Figura A.7: Matriz de similaridade da relação entre categorias e tópicos do Outubro Rosa do ano de 2015

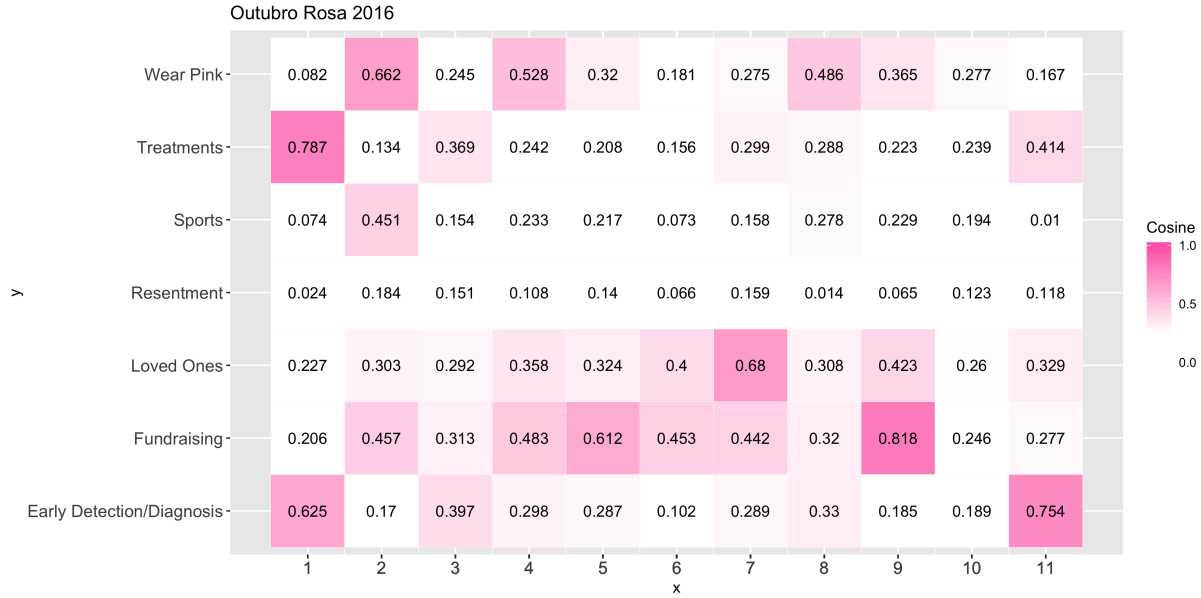


Figura A.8: Matriz de similaridade da relação entre categorias e tópicos do Outubro Rosa do ano de 2016

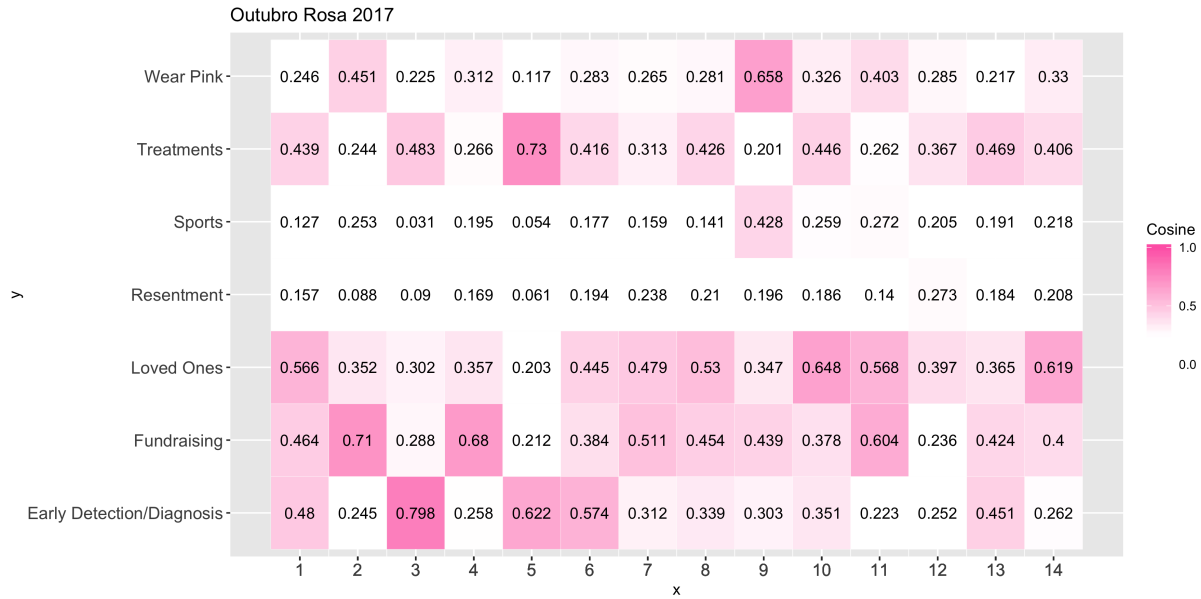


Figura A.9: Matriz de similaridade da relação entre categorias e tópicos do Outubro Rosa do ano de 2017


Figura A.10: Matriz de similaridade da relação entre categorias e tópicos do Outubro Rosa do ano de 2018

# ApêndiceB

## B.1 Wordclouds dos tópicos da campanha do Outubro Rosa

## B.2 Wordclouds dos tópicos da campanha do Novembro Azul

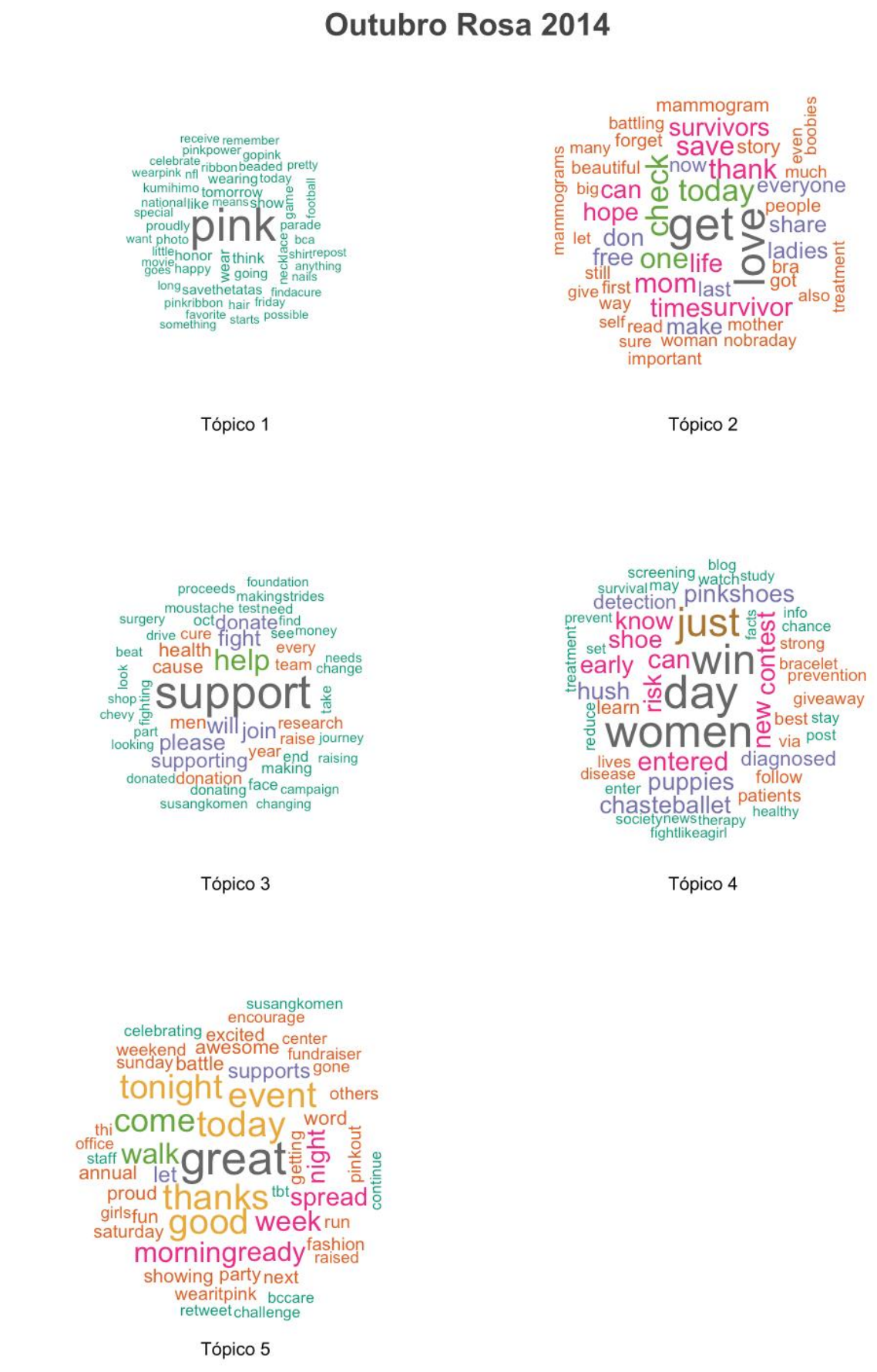


Figura B.1: Wordclouds dos tópicos da campanha do Outubro Rosa do ano de 2014

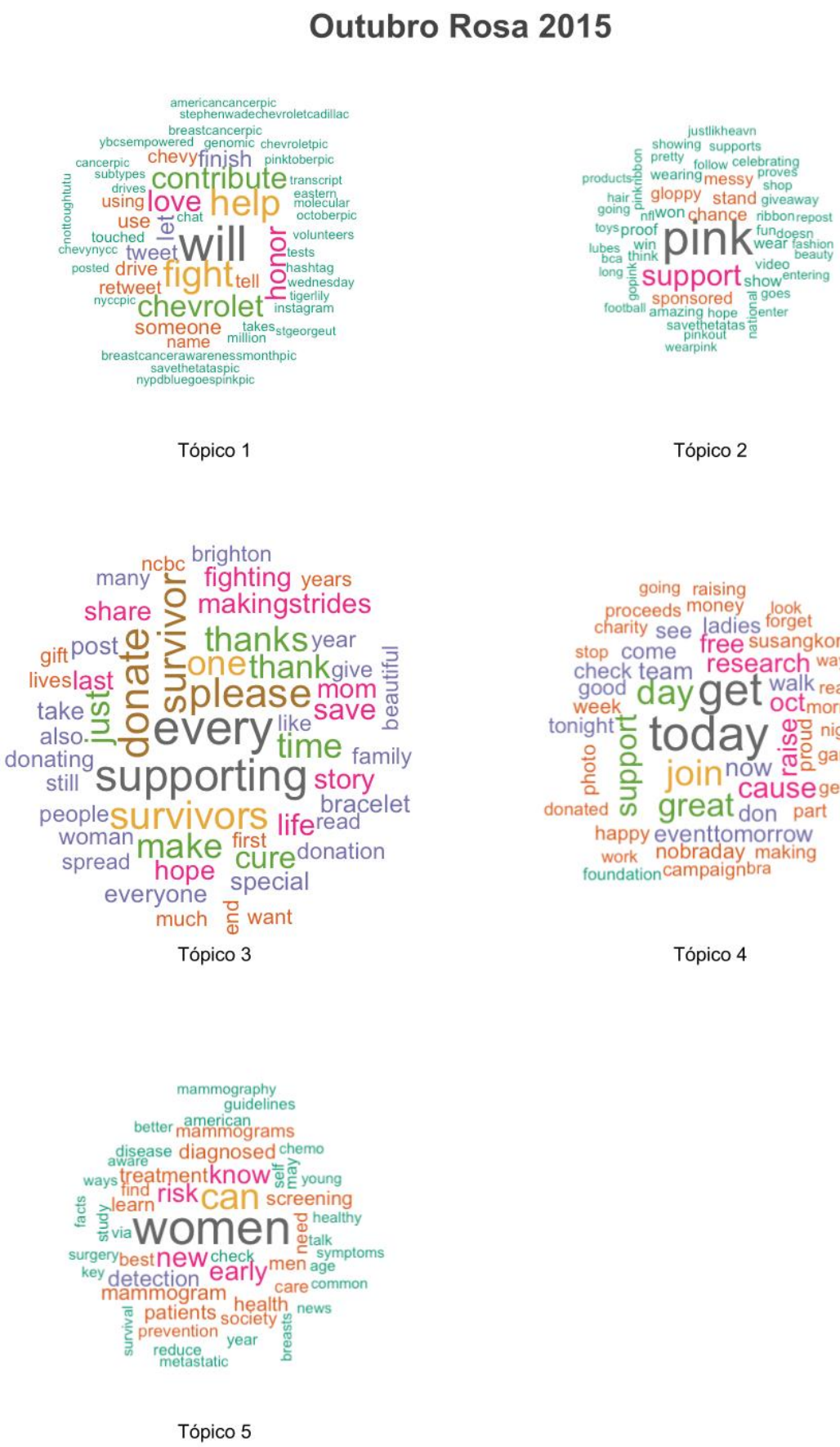


Figura B.2: Wordclouds dos tópicos da campanha do Outubro Rosa do ano de 2015

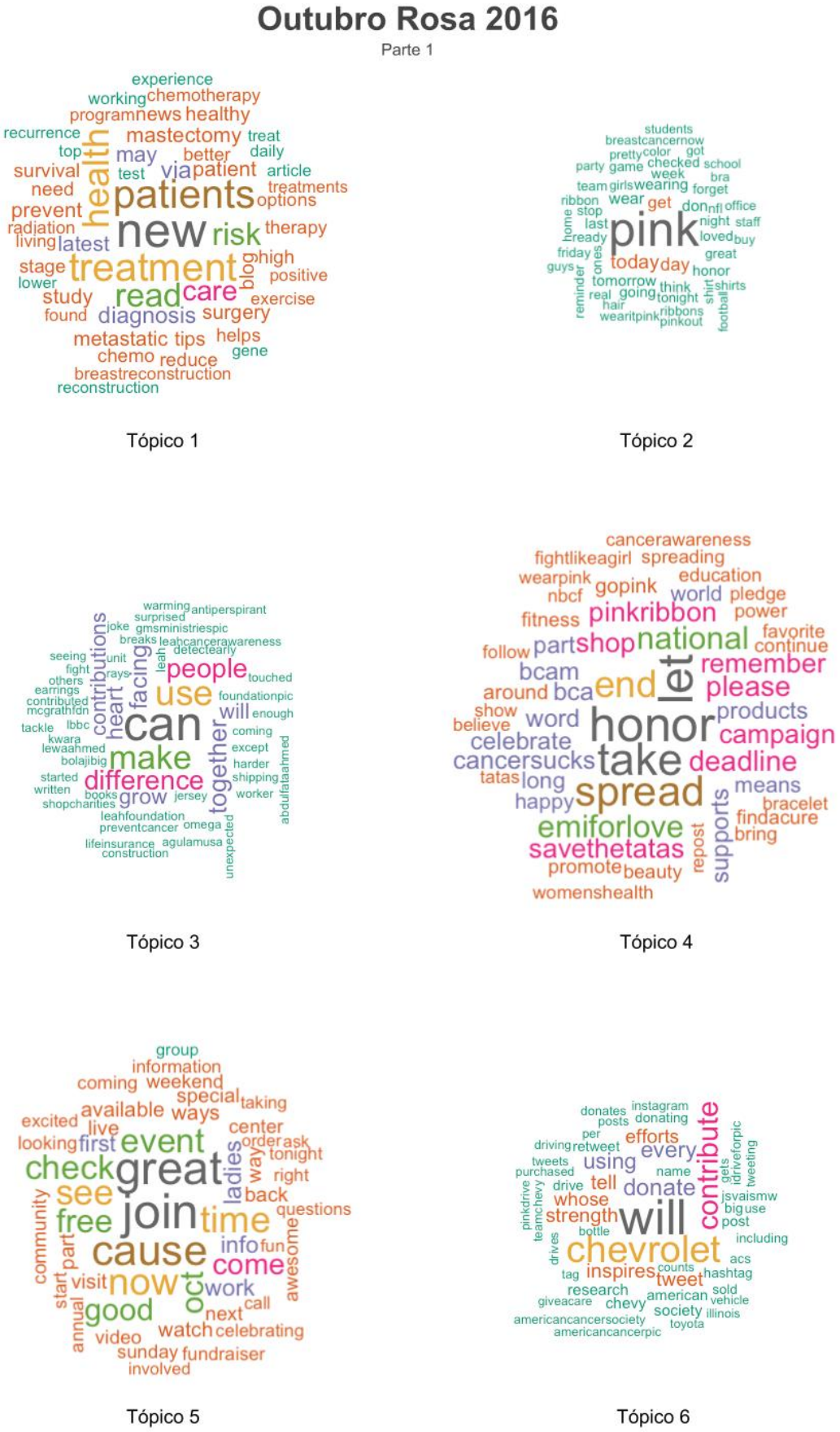


Figura B.3: Wordclouds dos tópicos da campanha do Outubro Rosa do ano de 2016 - Parte 1


Figura B.4: Wordclouds dos tópicos da campanha do Outubro Rosa do ano de 2016 - Parte 2

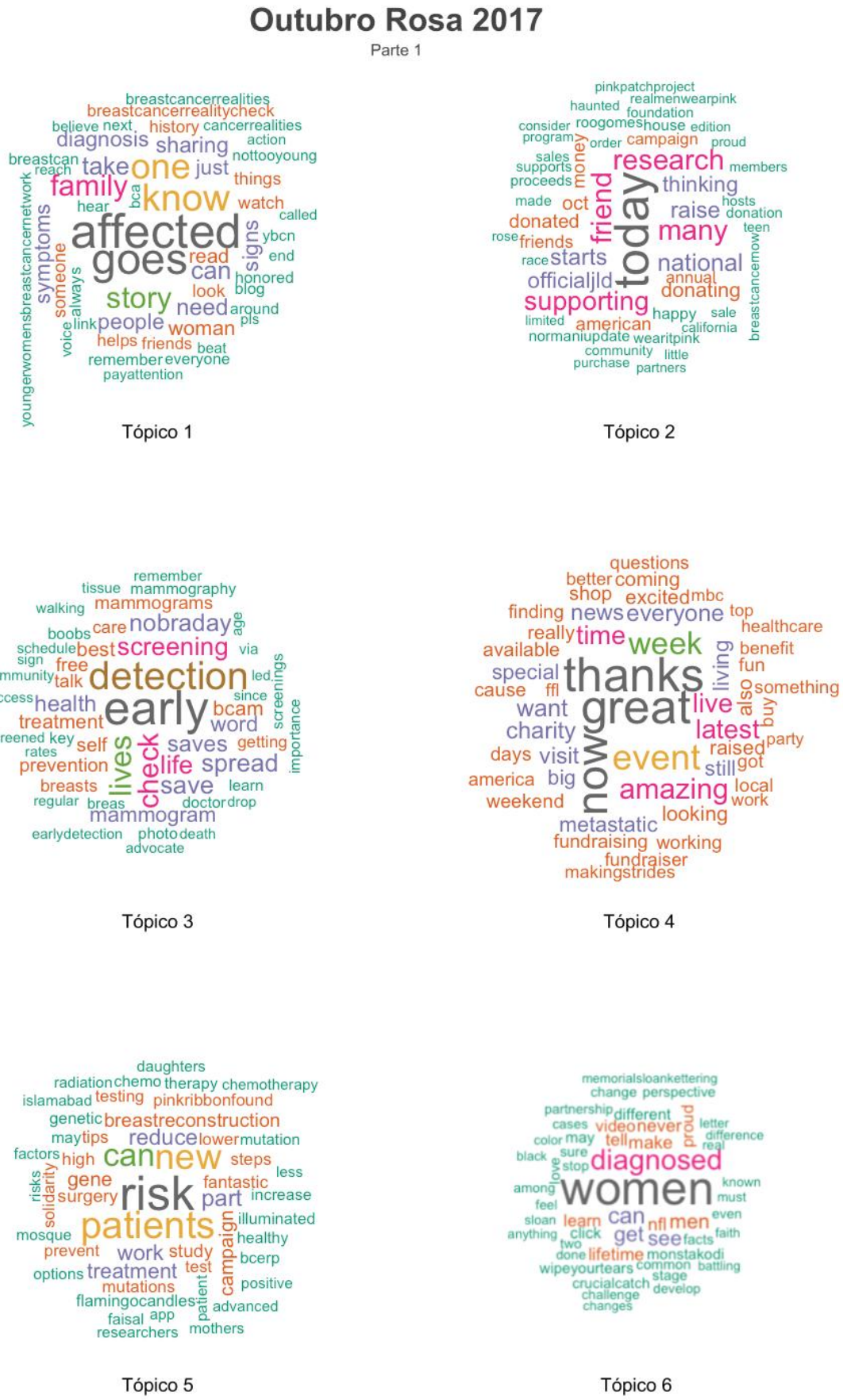


Figura B.5: Wordclouds dos tópicos da campanha do Outubro Rosa do ano de 2017 - Parte 1

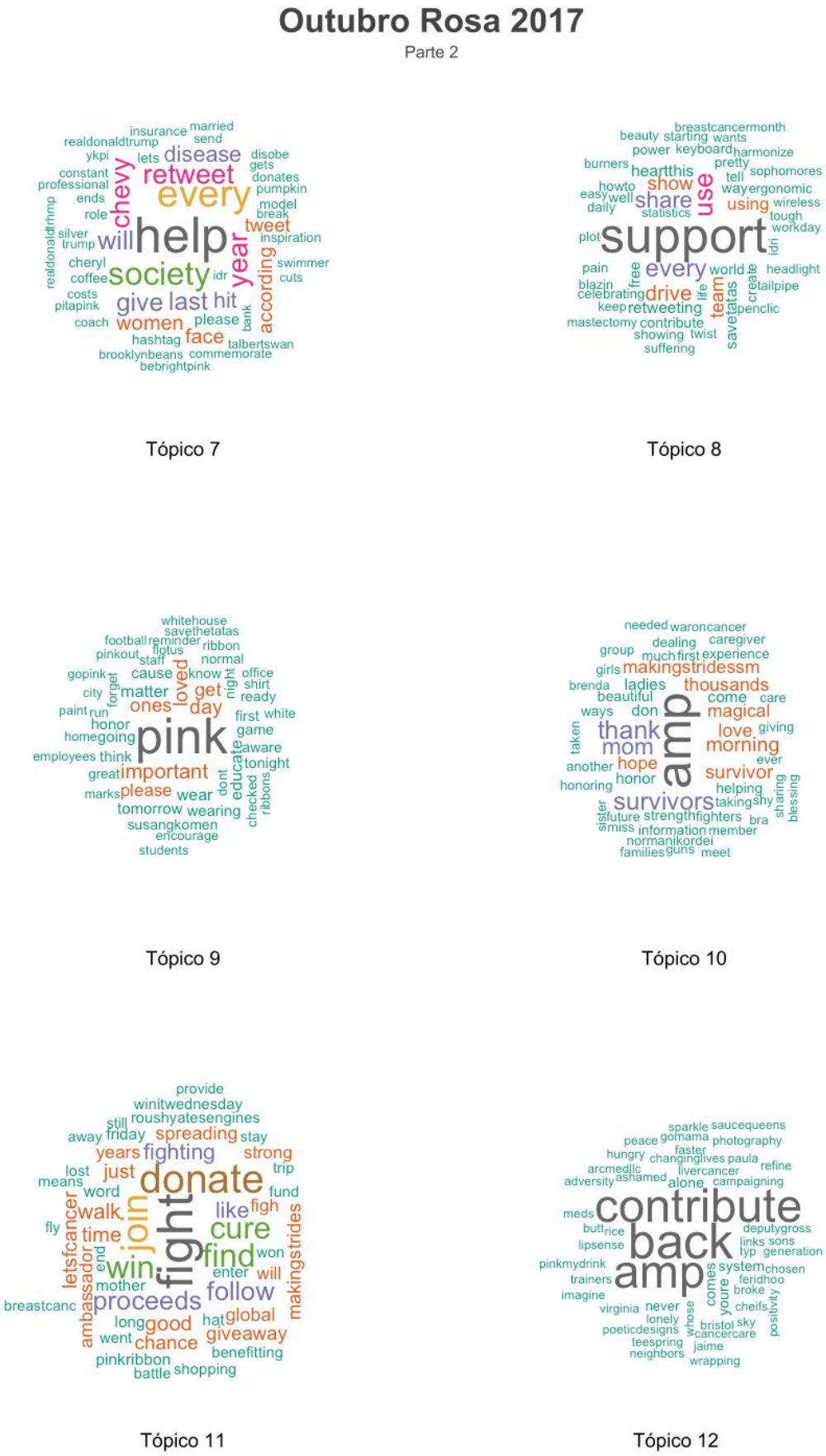


Figura B.6: Wordclouds dos tópicos da campanha do Outubro Rosa do ano de 2017 - Parte 2

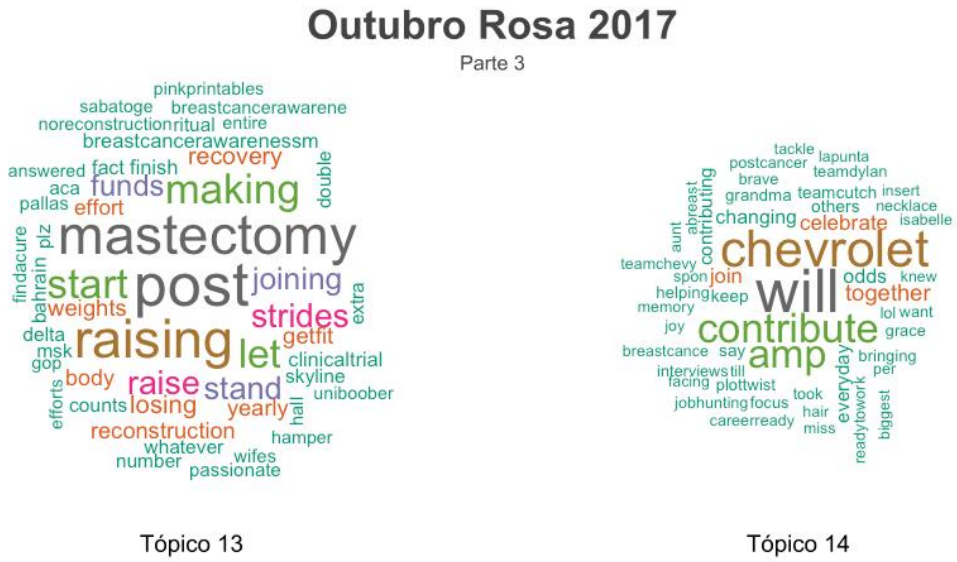


Figura B.7: Wordclouds dos tópicos da campanha do Outubro Rosa do ano de 2017 - Parte 3

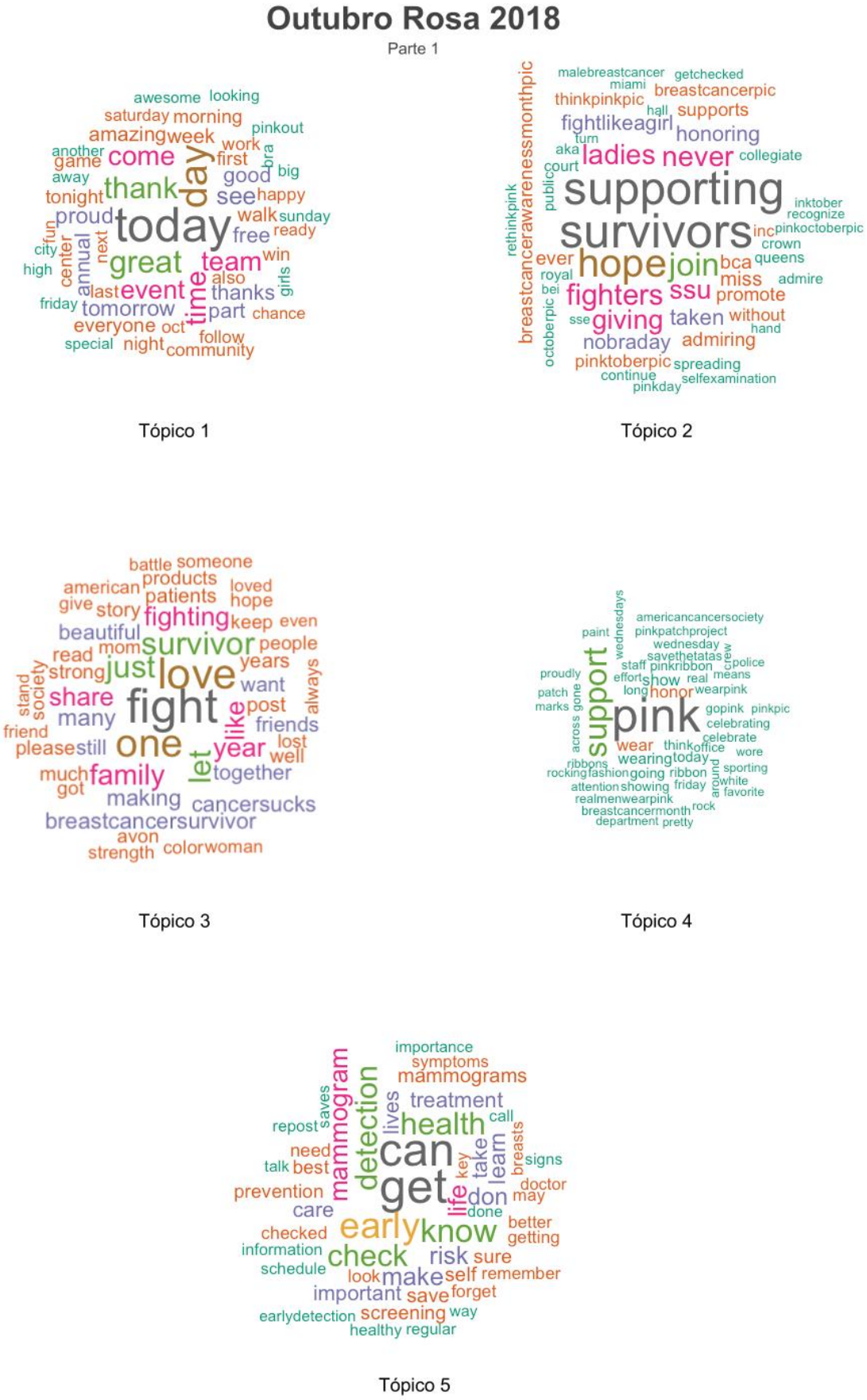


Figura B.8: Wordclouds dos tópicos da campanha do Outubro Rosa do ano de 2018 - Parte 1

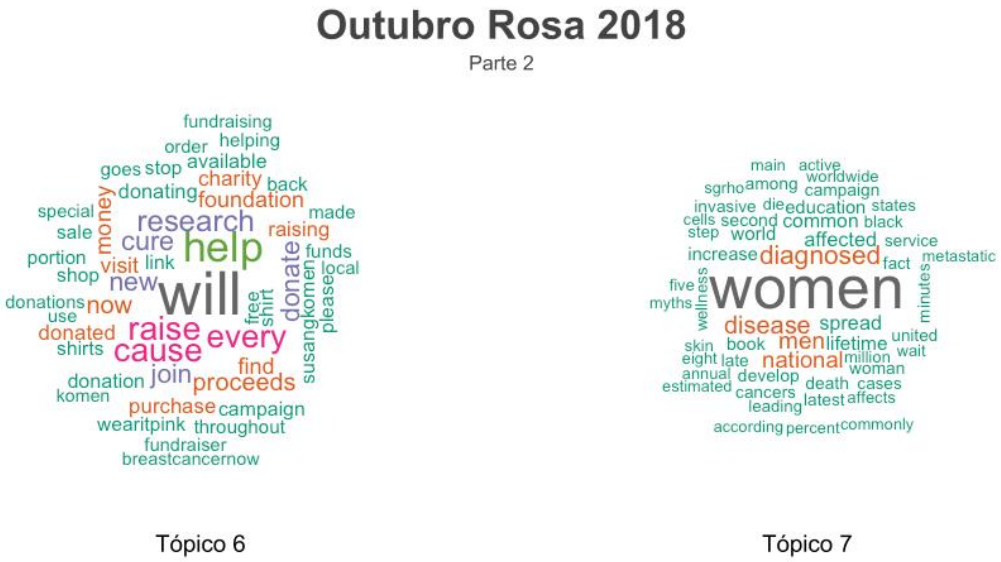


Figura B.9: Wordclouds dos tópicos da campanha do Outubro Rosa do ano de 2018 - Parte 2

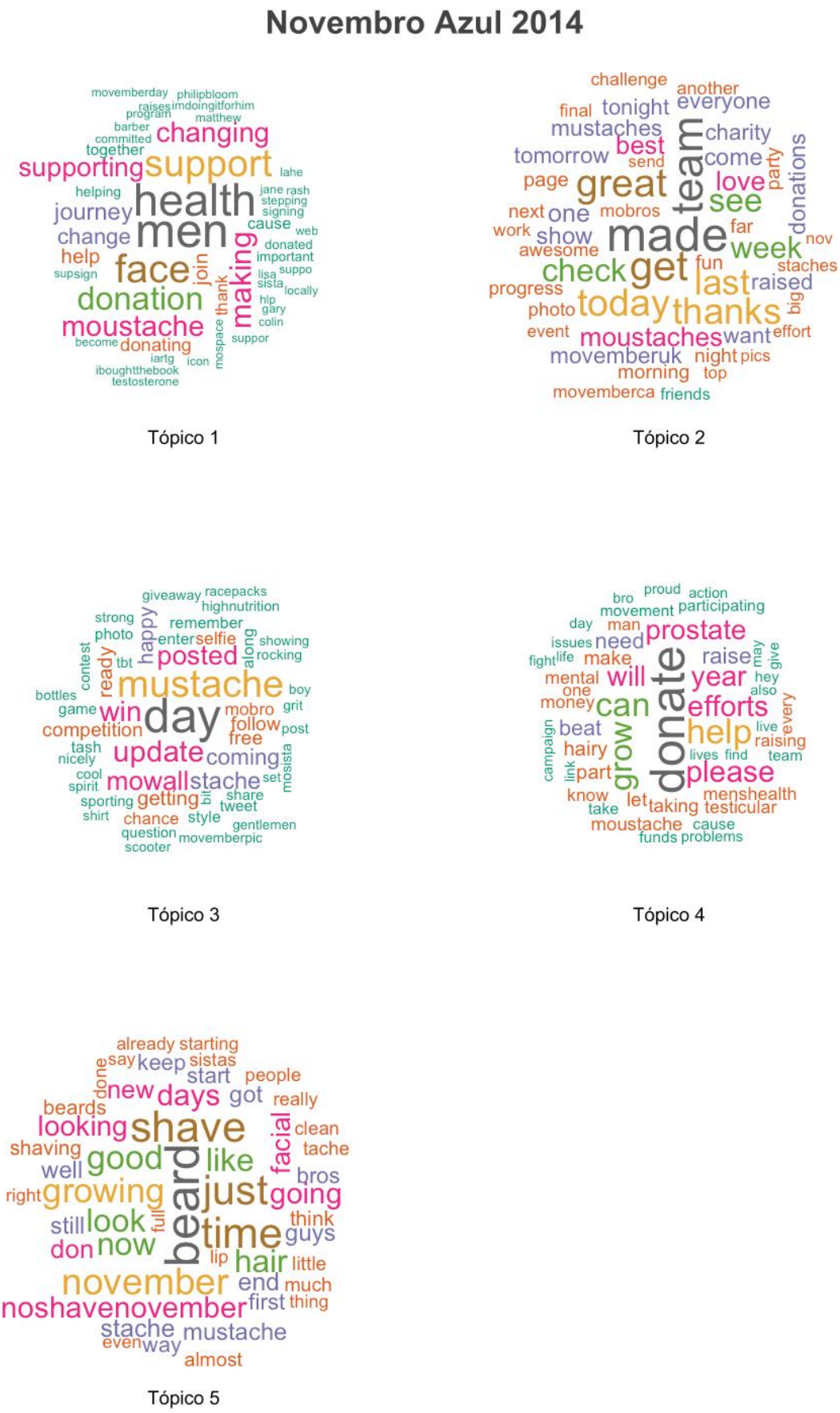


Figura B.10: Wordclouds dos tópicos da campanha do Novembro Azul do ano de 2014

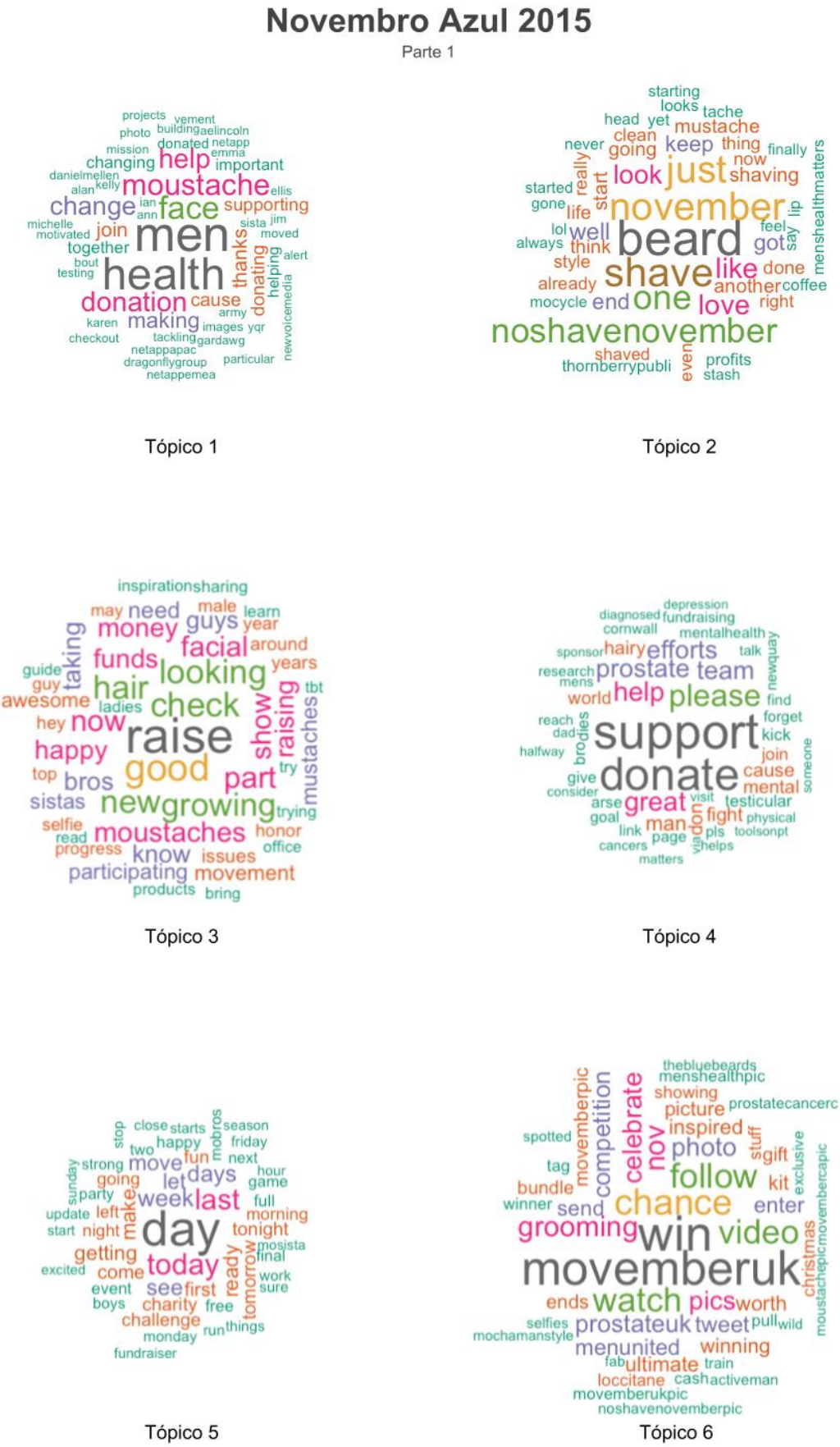


Figura B.11: Wordclouds dos tópicos da campanha do Novembro Azul do ano de 2015 - Parte 1

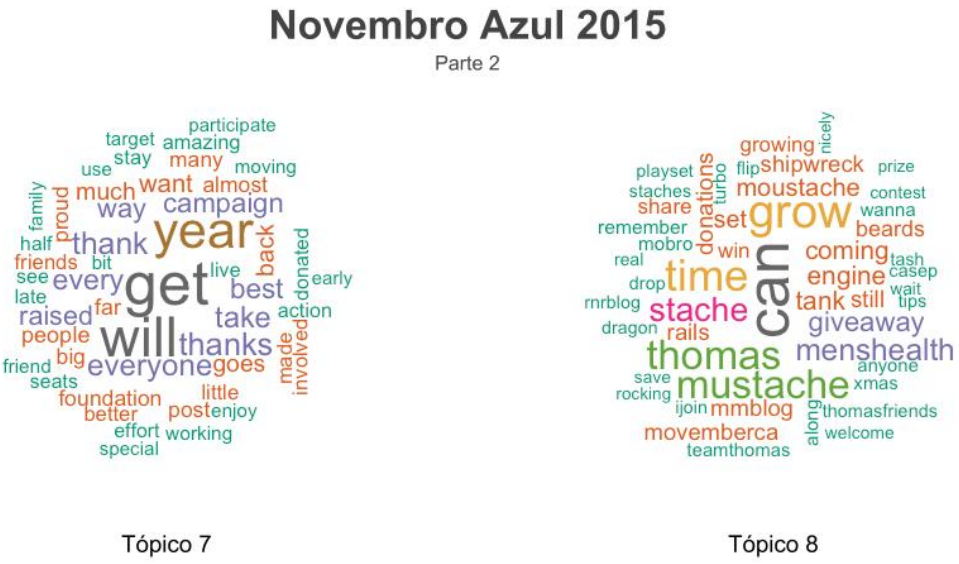


Figura B.12: Wordclouds dos tópicos da campanha do Novembro Azul do ano de 2015 - Parte 2

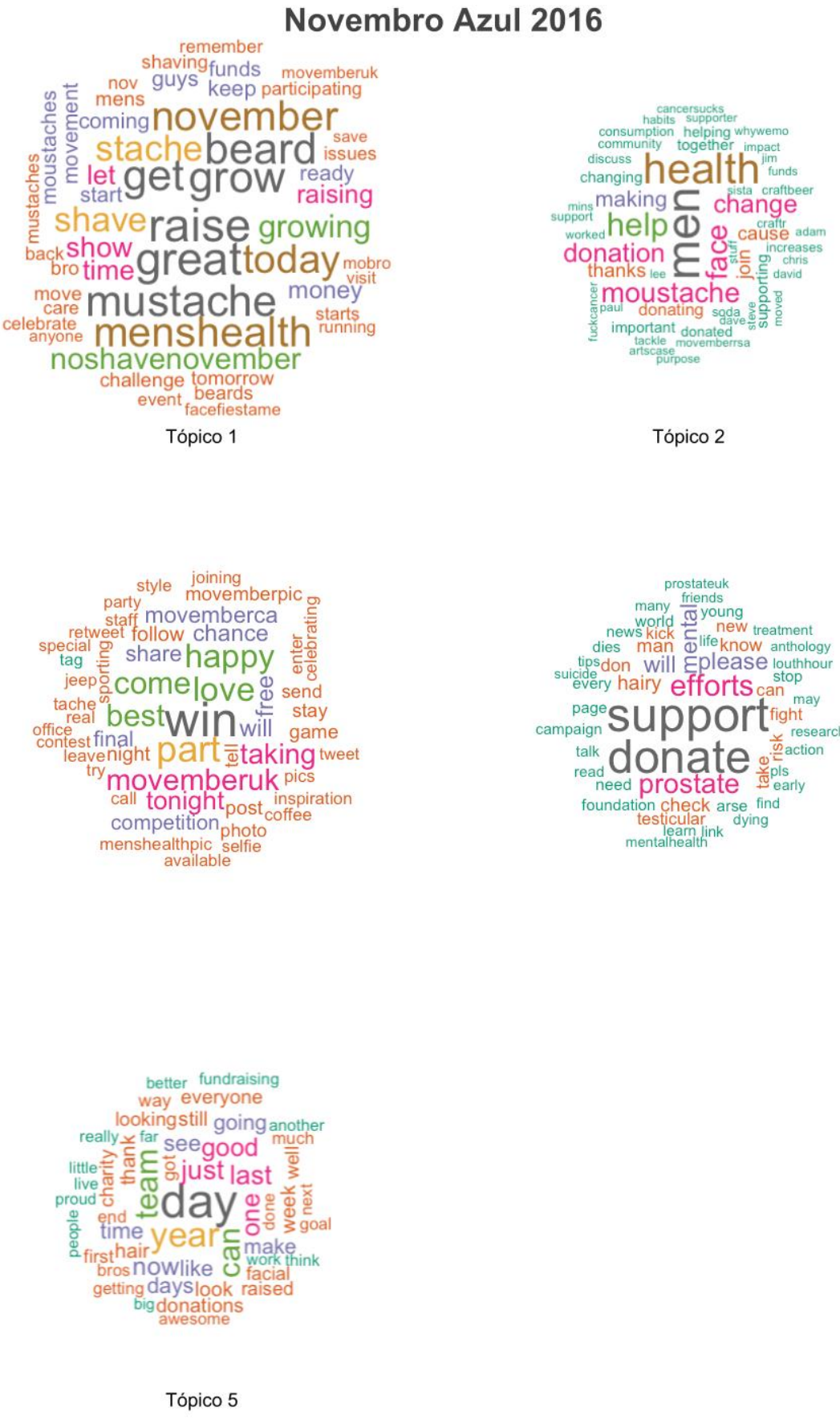


Figura B.13: Wordclouds dos tópicos da campanha do Novembro Azul do ano de 2016

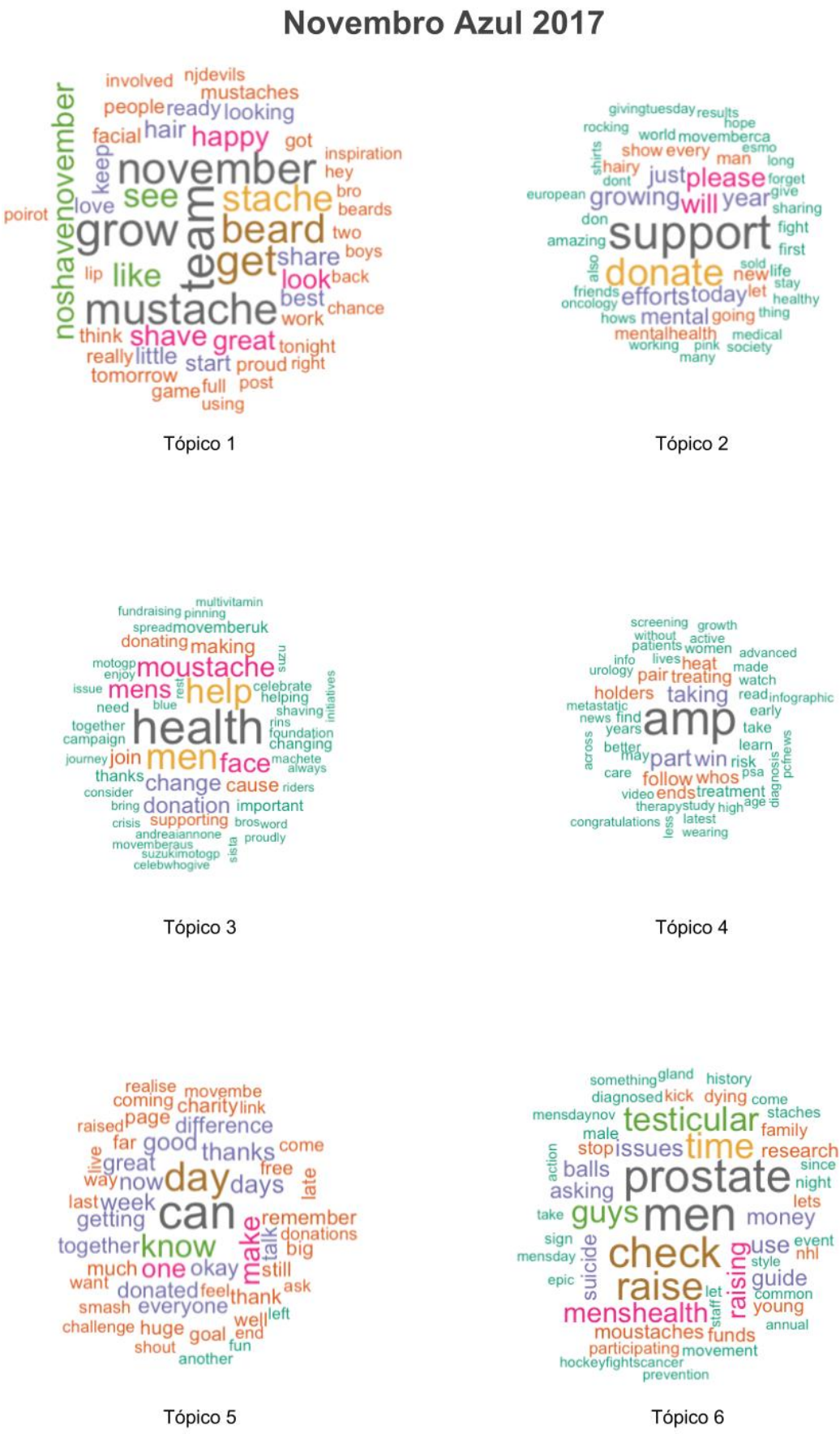


Figura B.14: Wordclouds dos tópicos da campanha do Novembro Azul do ano de 2017

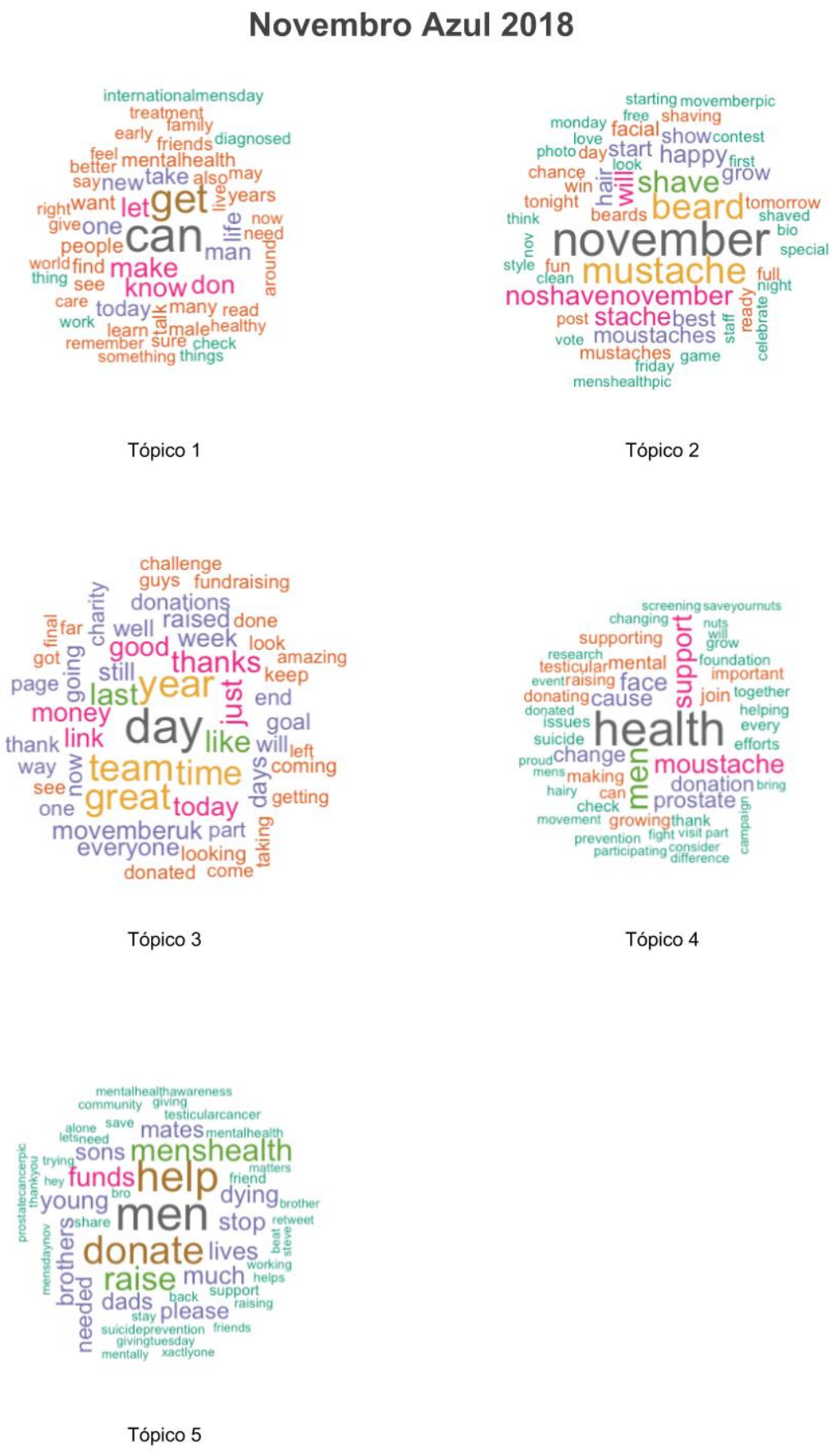


Figura B.15: Wordclouds dos tópicos da campanha do Novembro Azul do ano de 2018