Submission date: 25/02/2021
Resubmission date: 03/05/2021
Camera ready submission date: 08/07

1st round notification: 22/04/2021
2nd round notification: 24/06/2021
Available online: 20/08/2021
Publication date: 20/08/2021

Section: extended version
Extended versions of selected papers from the IX Brazilian Workshop on Social Network Analysis and Mining

# Análise e Previsão do Tom Emocional de Usuários em Comunidades de Saúde Mental no Reddit

# Title: Analysis and Prediction of Users' Emotional Tone in Reddit Mental Health Communities

**Bárbara Silveira, Fabricio Murai, Ana Paula Couto da Silva**

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

{barbarasilveira,ana.coutosilva,murai}@dcc.ufmg.br

***Abstract.*** *The rise in the number of people afflicted by mental health problems has placed these disorders among the main public health problems worldwide. As a result, user activity in communities related to mental health in online social networks has spiked. Here we characterize user activity in mental health-related communities on Reddit and analyze how user interactions through posts and comments influence their emotional state. In particular, we investigate whether seeking help on these networks results in changes in the feelings expressed by users over time. We observe that authors of negative posts often write rosier comments after engaging in discussions, indicating that users' emotional state can improve due to social support. Our results show that users who start discussions in these communities writing posts expressing negative feelings, tend to write more positive comments at the end, that is, they present an improvement in their emotional tone. In addition, we propose predictive models to capture the variation of the emotional tone of these users. Our models could assist in interventions promoted by health care professionals to provide support to the mentally-ill.*

***Keywords.*** *mental health; reddit; sentiment analysis; machine learning*

***Resumo.*** *O crescimento do número de pessoas atingidas por problema de saúde mental colocou tais distúrbios entre os principais problemas de saúde pública em todo o mundo. Como resultado, aumentou-se a procura por comunidades sobre saúde mental em redes sociais online. Neste artigo, nós caracterizamos a atividade de usuários em comunidades relacionadas à saúde mental no Reddit e analisamos como suas interações através de posts e comentários influenciam no seu estado emocional. Em particular, nós investigamos se a busca por auxílio nestas redes resulta em mudanças nos sentimentos expressos pelos usuários ao longo do tempo. Nossos resultados mostram que os usuários que iniciam discussões nestas comunidades com posts expressando sentimentos negativos, tendem a escrever comentários mais positivos ao final, indicando que o estado*

Cite as: Silveira, B., Murai, F., & Silva, A.P.C. (2021). Analysis and Prediction of Users' Emotional Tone in Reddit Mental Health Communities (Análise e Previsão do Tom Emocional de Usuários em Comunidades de Saúde Mental no Reddit). iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 14(2), 53-78. DOI: 10.5753/isys.2021.1877

*emocional dos mesmos pode ter melhorado em decorrência do suporte social provido por estas comunidades. Adicionalmente, propomos modelos preditivos para capturar a variação do tom emocional destes usuários. Nossos modelos poderiam auxiliar nas intervenções promovidas pelos profissionais de saúde para dar suporte aos indivíduos que sofrem de transtornos de saúde mental.*

***Palavras-Chave.*** *Saúde mental; Reddit; Análise de sentimentos; Aprendizado de máquina*

## 1. Introdução

Dados da Organização Mundial de Saúde (OMS) mostram que a cada 40 segundos uma pessoa morre por suicídio, sendo esta a segunda causa de morte entre jovens entre 15 e 29 anos[1]. Diversos fatores podem resultar nesta situação extrema, dentre eles, transtornos mentais, como a ansiedade, a depressão e a bipolaridade. Para citar algumas estatísticas, a ansiedade atinge 264 milhões de indivíduos no mundo. O transtorno bipolar[2] afeta cerca de 60 milhões de pessoas e, muitas vezes, este problema pode ser confundido com um caso de depressão ou de ansiedade, levando a um tratamento inadequado. Apesar desses números preocupantes, muitas das pessoas que sofrem de distúrbios mentais não recebem tratamento. Segundo a OMS, a cada 4 pessoas, 3 não recebem qualquer tipo de tratamento e 45% da população mundial vive em um país com menos de um psiquiatra para cada 100 mil habitantes[3].

Muitas pessoas que precisam de apoio acabam não sendo tratadas devido a ausência de recursos para apoiá-las ou pelo estigma social associado aos transtornos mentais [Barney et al. 2006], o que pode gerar consequências irreversíveis. Assim, diferentes ferramentas para auxiliar pessoas que passam por problemas relacionados à saúde mental estão sendo exploradas nos últimos anos. Em particular, podemos ressaltar o papel das redes sociais online. Inicialmente focadas em fomentar amizades, trocar imagens ou vídeos, passaram a conectar pessoas dispostas a compartilhar pensamentos, sentimentos e experiências, contribuindo para a melhoria do bem-estar dos indivíduos que sofrem desses transtornos [De Choudhury 2013].

Neste contexto, o presente artigo analisa como as interações através de *posts* e comentários influenciam o estado emocional dos usuários de comunidades de saúde mental no Reddit[4], representado pelo tom emocional de suas mensagens. O Reddit é um site de fóruns com características de redes sociais: é composto por comunidades (subreddits), onde os usuários compartilham suas experiências e dúvidas sobre os mais diversos assuntos. Nesta rede, um usuário inicia uma *thread* ao publicar um *post*. Este *post* pode ser respondido por comentários de outros usuários ou mesmo daquele que escreveu o *post*. Comentários, por sua vez, podem ser respondidos por outros comentários. Dados de 2019

[1] https://www.who.int/docs/default-source/mental-health/suicide/live-life-brochure.pdf?sfvrsn=6ea28a12_2&download=true

[2] https://www.who.int/en/news-room/fact-sheets/detail/mental-disorders

[3] https://www.who.int/mental_health/evidence/atlas/interactive_infographic_2015.pdf

[4] https://www.reddit.com/

mostram que o Reddit é constituído por 430 milhões de usuários ativos por mês. Ao todo são 199 milhões de *posts* e 1,7 bilhões de comentários[5].

Nosso trabalho tem três objetivos principais: (1) determinar os padrões de atividade dos usuários dessas comunidades de saúde mental, (2) compreender como elas são utilizadas para apoiar na melhoria do tom emocional de seus participantes e (3) definir modelos que permitam acompanhar a evolução do tom emocional do usuário.

A partir dos objetivos listados, nossas contribuições são:

1. *Caraterização dos padrões de atividade dos usuários e de suas interações.* Apresentamos uma caracterização dos principais subreddits relacionados à saúde mental, que inclui uma análise de como as atividades de um usuário estão distribuídas no tempo. Identificamos perfis de usuários de acordo com o tipo e o volume de atividades dentro da comunidade. Tais perfis podem ser úteis para identificar a abordagem mais efetiva para ajudar um usuário. Analisamos também as interações entre usuários, que se dão por meio de posts e comentários, para entender a estrutura do grafo social subjacente. Concluímos que, diferente das redes sociais tradicionais, os subreddits se organizam principalmente em torno do conteúdo e não dos relacionamentos.
2. *Análise do Tom Emocional dos Usuários.* Tom emocional (TE) é o sentimento (positivo e negativo combinados) extraído através de um texto. Sabendo que existem usuários que procuram estas comunidades em busca de ajuda e outros que estão dispostos a ajudar, analisamos a variação do estado emocional de um usuário dentro de uma *thread*, verificando se os outros usuários influenciam no TE do usuário que iniciou a discussão. Em geral, observamos que o TE do último comentário do autor da *thread* é maior que o do *post* que a iniciou e também superior à média do TE dos comentários da árvore de discussão.
3. *Modelos para Previsão do Tom Emocional dos Usuários.* A partir das análises do TE dos usuários, propomos modelos para prever a evolução do TE dos participantes destas comunidades. Utilizamos o *Multilayer Perceptron* que consiste em uma rede neural com pelo menos três camadas. Estes modelos capturam com boa acurácia (MSE $0.66$) a variação do estado emocional dos usuários (que assume valores entre $-2$ e $2$). Uma possível aplicação dos modelos é auxiliar intervenções promovidas por profissionais da área de saúde em redes sociais.

Este artigo é um desdobramento do nosso trabalho anterior [Silveira et al. 2020] no qual analisamos a influência das interações dentro das comunidades do Reddit relacionadas à saúde mental na variação do tom emocional dos usuários e propomos modelos de previsão para capturar sua evolução. Aqui estendemos o artigo anterior através de uma detalhada caracterização dessas comunidades (Seção 3), dos padrões de atividade dos usuários (Seção 5.1) e de suas interações (Seção 5.2).

O restante do texto está organizado da seguinte forma. A Seção 2 descreve os principais trabalhos relacionados e a Seção 3 apresenta os dados utilizados e uma caracterização geral das atividades dos usuários. A Seção 4 detalha os métodos utilizados. Os resultados da clusterização dos usuários segundo o padrão de atividades, da análise do

[5]`https://redditblog.com/2019/12/04/reddits-2019-year-in-review/`

tom emocional e dos modelos aplicados são apresentados na Seção 5. Finalizando, as implicações deste trabalho e trabalhos futuros são discutidos na Seção 6.

## 2. Trabalhos Relacionados

Com o passar dos anos, as Redes Sociais Online (RSO) passaram a permitir o compartilhamento de vários tipos de mídia, resultando em um grande número de trabalhos na literatura que utilizam dados oriundos destas redes para analisar o comportamento dos usuários em diferentes aspectos da vida cotidiana. Por exemplo, podemos citar o estudo realizado por [Cunha et al. 2017] que propõe um *framework* para relevar relações de causalidade entre interações de usuários no subreddit "loseit" e perda de peso.

Além da saúde física, alguns trabalhos na literatura utilizam RSOs como ferramenta para entender problemas relacionados à saúde mental. Os autores em [Blair and Abdullah 2018] utilizam dados do Instagram para identificar e analisar os desafios enfrentados pelos usuários que divulgam seus problemas de saúde mental nesta rede. O trabalho mostra que a mídia social pode ter um impacto positivo para esses usuários, uma vez que permite que eles construam suas próprias comunidades. O trabalho feito por [Islam et al. 2018] utiliza o Facebook como objeto de estudo para detecção de depressão através de técnicas de aprendizado de máquina. Os autores mostram que o método proposto, que utiliza características psicolinguísticas, é eficiente para realizar a classificação dos usuários com depressão. Em outro trabalho, os autores em [Weerasinghe et al. 2019] analisaram os tópicos no Twitter para descobrir padrões de linguagem que diferenciam indivíduos com doenças mentais de um grupo de controle. Como resultado, os pesquisadores confirmaram certos padrões e descobriram outros novos.

Ainda neste contexto, outros trabalhos abordam a detecção da ansiedade, bipolaridade, depressão e suicídio em diferentes redes sociais e procuram entender o comportamento dos usuários presentes nessas redes sociais como forma de propor políticas que contribuam para a diminuição do volume de pessoas afetadas por estes problemas [Shen and Rudzicz 2017, Wolohan et al. 2018, Wongkoblap et al. 2019, Gruda and Hasan 2019, Sahota and Sankar 2019, Baba et al. 2019, Silveira et al. 2018]. Considerando o Reddit, os autores em [De Choudhury and De 2014] utilizam dados coletados desta rede para analisar os posts e comentários dos usuários, investigando como o grau de desinibição nos comentários e *posts* feitos por usuários anônimos se difere daqueles feitos pelos usuários que se identificam. Em [Silveira Fraga et al. 2018], apresentamos os principais tópicos discutidos pelos participantes dos subreditts relacionados à saúde mental, além de apresentar a caracterização (formato e tamanho) das *threads* criadas pelos usuários.

A maior parte dos trabalhos listados anteriormente tem como foco analisar o comportamento das pessoas que sofrem de doenças relacionadas à saúde mental, visando auxiliar na elaboração de políticas de saúde pública, desenvolver aplicações que auxiliem estas pessoas, entre outras medidas. Diferentemente dos trabalhos listados, neste artigo analisamos o tom emocional de usuários de comunidades online relacionadas a transtornos mentais com o objetivo de mensurar o impacto que as interações nessas comunidades geram no estado de seus usuários. Nosso estudo considera o comportamento do usuário dentro de uma *thread*, desde o momento em que publica um *post* até o momento em que

realiza seu último comentário e, através de diferentes modelos baseados em redes neurais, prevemos as mudanças no seu tom emocional no intervalo no qual o mesmo interage com outros usuários pertencentes à essas comunidades.

Por fim, ressaltamos alguns trabalhos encontrados na literatura que, apesar de não possuírem foco na análise de comunidades que discutem distúrbios de saúde mental, visam analisar a variação do "humor" dos usuários ao interagirem com seus pares em redes sociais online. Os autores em [Ballona et al. 2015] estudam a influência (positiva ou negativa) dos posts do Papa no humor dos seus seguidores no Twitter. O artigo [Pellert et al. 2020] apresenta um estudo sobre a dinâmica individual do afeto que captura as mudanças dos estados emocionais ao longo do tempo, independentemente da interação social ou de outros estímulos externos. Esses dois estudos avaliam a variação dos estados emocionais, resultantes das interações com seus pares, utilizando diferentes metodologias aplicadas no nosso trabalho e, principalmente, em comunidades com características muito distintas das comunidades que discutem temas relacionados à saúde mental.

## 3. Base de dados e caracterização das atividades dos usuários

Existem mais de 25 subreddits que focam na discussão de transtornos mentais. Destes, os quatro com o maior número de *posts* e comentários [Gkotsis et al. 2016] são[6]: Depression (/r/depression), SuicideWatch (/r/suicide), Anxiety (/r/Anxiety) e Bipolar (/r/bipolar).[7] Neste trabalho, utilizamos todos os *posts* e comentários compartilhados nestas comunidades entre Janeiro de 2011 e Dezembro de 2017.[8]

**Estatísticas gerais dos subreddits.** Como passo inicial do nosso estudo, realizamos uma caracterização geral dos dados que inclui o cálculo de estatísticas relacionadas ao volume de usuários ativos, ao volume de posts e comentários, assim como a atributos desses diferentes tipos de mensagem. Dentre os atributos, analisamos o comprimento (medido em número de palavras) e o "score" dos *posts* e comentários escritos por usuários de cada comunidade. O "score", também conhecido como "Karma", corresponde a quantidade de votos recebidos (podendo ser negativos ou positivos) pelos usuários.

A Tabela 1 apresenta estatísticas básicas agregadas sobre o período selecionado. Consideramos todos *posts* e comentários, inclusive aqueles que foram removidos, deletados ou estavam vazios. Nas estatísticas apresentadas, desconsideramos os *posts* e comentários realizados por usuários deletados (i.e., usuários que foram excluídos do Reddit), porque não é possível distinguir os autores destas publicações. Podemos notar que a comunidade Bipolar é a que realiza, proporcionalmente, mais *posts* e comentários. Além disso, a média dos comentários por usuário nesta comunidade é cerca de três vezes maior que nas outras três. Isto é um indício que os usuários nesta comunidade são mais engajados durante as discussões.

[6]Não consideramos o subreddit Opiates (/r/opiates) por estar mais ligado à dependência de remédios.

[7]Subreddits em português: Depressão, Risco de suicídio, Ansiedade e Bipolar. A notação /r/<nome> é utilizada pelo Reddit para referenciar um determinado subreddit.

[8]Os dados foram recuperados a partir de: `http://files.pushshift.io/reddit/`. Os dados pós-processados e utilizados em nossas análises estão disponíveis em `https://doi.org/10.5281/zenodo.4266616`.

| | Depression | SuicideWatch | Anxiety | Bipolar | Total |
|---|---|---|---|---|---|
| Usuários Únicos | 333.624 | 162.363 | 121.945 | 35.081 | 569.122 |
| Posts | 468.507 | 169.541 | 142.168 | 73.254 | 853.470 |
| Comentários | 2.128.991 | 1.024.171 | 685.867 | 607.262 | 4.446.291 |
| Posts/usuário | 1,40 | 1,04 | 1,17 | 2,09 | 1,51 |
| Comentários/usuário | 6,38 | 6,31 | 5,62 | 17,31 | 7,81 |
| Comentários/post | 4,54 | 6,04 | 4,82 | 8,29 | 5,21 |

Tabela 1. Estatísticas básicas de cada comunidade.

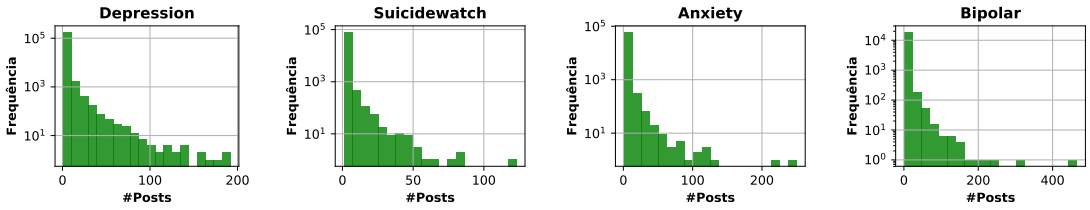



Figura 1. Número de *Posts* por usuário.

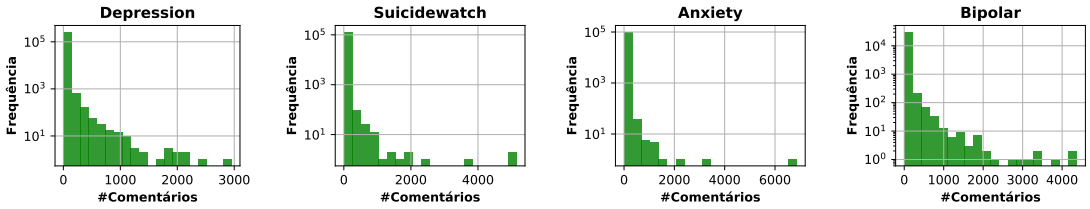



Figura 2. Número de Comentários por usuário.

Analisamos também a distribuição do número de publicações por usuário. As Figuras 1 e 2 apresentam, respectivamente, o histograma do número de *posts* e de comentários por usuário em cada comunidade. Observa-se que ambas as distribuições têm cauda pesada, porém o número de comentários por usuário pode chegar a alguns milhares, enquanto o número de posts ficou limitado a algumas centenas.

A Tabela 2 apresenta a média e a mediana dos atributos score e número de palavras calculadas sobre posts, comentários e publicações em geral, em cada subreddit. Observamos que o score de um *post* é, em média, 3,9 vezes maior que o score de um comentário. Uma possível interpretação é que os autores dos *posts* recebem mais suporte social do que os autores dos comentários. Outra interpretação é que os *posts* recebem mais votos positivos porque naturalmente atraem mais atenção que os comentários associados a eles. Além disso, note que os valores são, na média, sempre positivos, indicando que os usuários incentivam uns aos outros com votos positivos, ao invés de inibi-los com votos negativos. Apesar da média do score do *post* ser acima de 5, em todas comunidades, mais de 50% dos *posts* tem score bem menor que média. Isto mostra que existe grande variação entre os scores nos *posts* das comunidades.

| | **Depression** | | **SuicideWatch** | | **Anxiety** | | **Bipolar** | |
|---|---|---|---|---|---|---|---|---|
| | $\bar{X}$ | $X_{50}$ | $\bar{X}$ | $X_{50}$ | $\bar{X}$ | $X_{50}$ | $\bar{X}$ | $X_{50}$ |
| Score Post | 8,25 | 2 | 5,06 | 3 | 9,22 | 2 | 10,59 | 4 |
| Score Comentário | 2,39 | 1 | 1,52 | 1 | 2,35 | 1 | 2,13 | 2 |
| Score Geral | 3,45 | 1 | 2,02 | 1 | 3,53 | 2 | 3,04 | 2 |
| Palavras Post | 181,20 | 99 | 188,86 | 105 | 162,61 | 105 | 136,94 | 79 |
| Palavras Comentário | 54,33 | 29 | 55,36 | 27 | 65,71 | 40 | 55,73 | 32 |
| Palavras Geral | 77,21 | 33 | 74,32 | 30 | 82,34 | 44 | 64,47 | 34 |

Tabela 2. Score e Tamanho das Palavras nas Comunidades ($\bar{X}$: média, $X_{50}$: mediana)

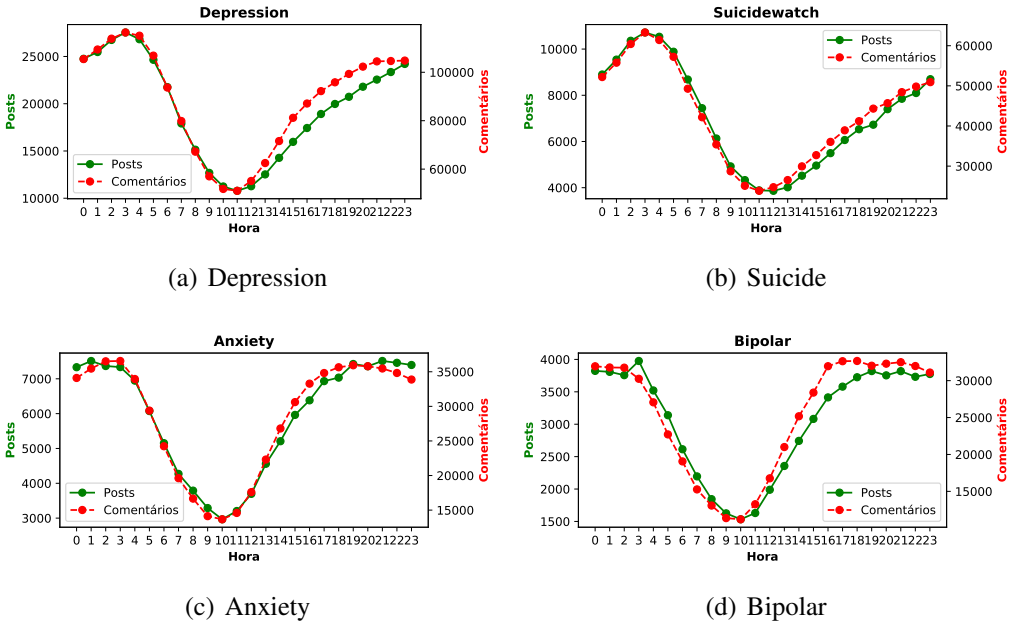
Em relação ao tamanho médio do texto (em quantidade de palavras), observamos que os *posts* são aproximadamente 2 vezes maiores que os comentários. Além disso, 50% dos *posts* possuem pelo menos 99 palavras no Depression, 105 no SuicideWatch, 105 no Anxiety e 79 no Bipolar. Este número é 54,6% do tamanho médio dos *posts* no Depression, 55,6% no SuicideWatch, 64,6% Anxiety e 57,7% no Bipolar. A razão entre a média e a mediana do tamanho dos comentários para cada comunidade é bem próxima a razão calculada para os *posts*. Em suma, há grande variabilidade no tamanho dos *posts* e comentários em todas comunidades.

**Distribuição de atividades de um usuário no tempo.** Conduzimos uma série de análises para melhor entender como os usuários realizam suas atividades (posts e comentários) em relação à várias escalas de tempo. Mais precisamente, estudamos como as atividades estão distribuídas ao longo das horas do dia e dos dias da semana. Calculamos também a distribuição de tempo entre duas atividades consecutivas e a distribuição do *lifespan* dos usuários em cada comunidade (i.e., intervalo entre primeira e última atividade).

A Figura 3 apresenta a variação do volume de atividades em cada comunidade, em diferentes horas do dia, segundo o fuso UTC, que foi o fuso adotado para registrar os timestamps pela base que utilizamos. Uma vez que a maioria dos usuários do Reddit, no período da coleta de dados, são originários dos Estados Unidos, utilizamos o fuso horário do centro médio da população americana[9] (UTC-5 para timestamp das atividades) para fazer uma análise, mesmo que aproximada, da distribuição das atividades dos usuários ao longo das horas do dia e dos dias da semana. Esta heurística pode gerar localização geográfica incorreta, com impactos na distribuição de atividades ao longo do tempo. No entanto, os dados disponibilizados não fornecem outros mecanismos para a determinação da localização geográfica real do usuário.

Neste fuso, podemos afirmar que o pico no volume de atividades nestas comunidades ocorre entre 20:00 e 22:00 e que o menor volume de atividades ocorre de madrugada, logo antes das 6:00. Além disso, a razão entre o número máximo e o mínimo de *posts* varia entre 2,5 (Anxiety) e o 2,8 (SuicideWatch) ao longo do dia. Já a mesma razão calculada com base nos comentários varia entre 2,3 (Depression) e 2,9 (Bipolar).

[9] `https://en.wikipedia.org/wiki/Mean_center_of_the_United_States_population`




(a) Depression (b) Suicide

(c) Anxiety (d) Bipolar

Figura 3. Atividades dos usuários (total de posts e comentários) em cada hora do dia.

A Figura 4 apresenta as atividades dos usuários em cada dia da semana.[10] Os maiores níveis de atividades ocorrem às segundas-feiras, com exceção do subreddit Bipolar, cujo pico acontece na quarta-feira. Em contrapartida, os menores níveis de atividades ocorrem aos sábados nos 4 subreddits. A diferença de postagens entre a segunda-feira para o sábado é 18,2% no Depression, 15,9% no SuicideWatch, 23,9% no Anxiety e 18,2% no Bipolar.

Como esperado, o histograma do dia de criação dos comentários segue o mesmo comportamento dos *posts*. O decaimento, considerando as segundas-feiras e os sábados, é 16,4% no Depression, 14,1% no SuicideWatch, 19,4% no Anxiety e 14,1% no Bipolar. Os valores em cada comunidade, em termos percentuais, são muito semelhantes. Também observamos que a queda nos *posts* é sempre maior que a nos comentários. A comunidade que mais se difere das outras é a Anxiety que tem maior queda, tanto nos *posts* quanto nos comentários.

É interessante notar que as ocorrências de picos de atividades às segundas-feiras seguem o efeito "Blue Monday" apresentado por [Stone et al. 1985], que reporta que as segundas-feiras são frequentemente associadas a restrições, tais como lazer e felicidade. Por outro lado, a pesquisa apresentada por [Cranford et al. 2006] mostra que, geralmente, o humor negativo diminui e a energia aumenta durante os finais de semana, o que pode justificar a queda acentuada de atividades nessas comunidades aos sábados e domingos.

Adicionalmente analisamos os intervalos entre atividades de um mesmo usuário. A Figura 5 apresenta a CCDF (Função de Distribuição Cumulativa Complementar) da

[10]Como o objetivo da análise é mostrar a variação do volume de atividades durante o dia, acreditamos que a adoção de uma escala única (ou uma iniciando em zero) para todos os subreddits poderia mascarar tais oscilações, principalmente para os subreddits menores (e.g., Bipolar).

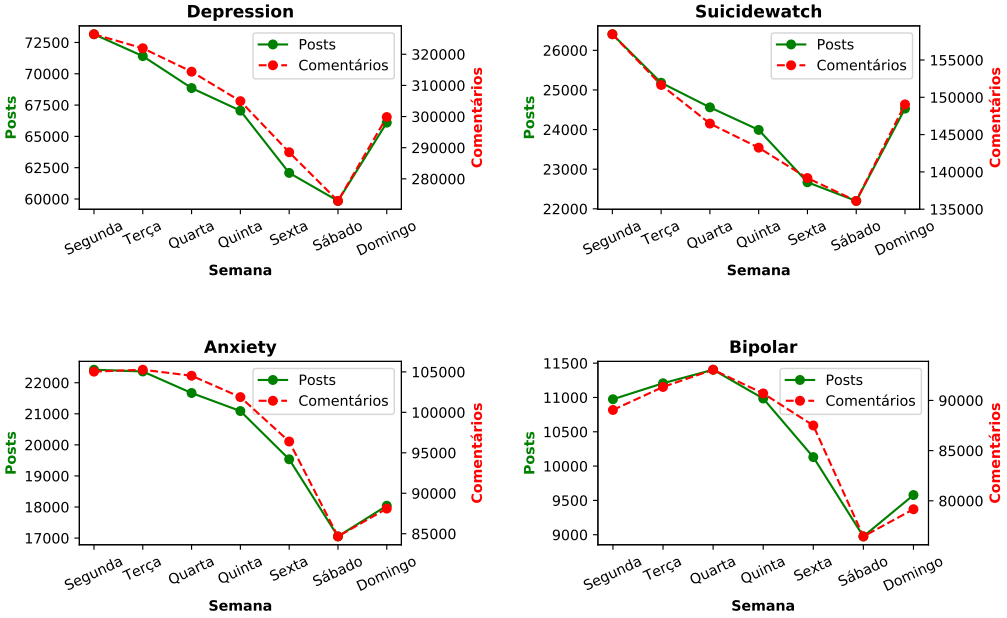



Figura 4. Atividades dos usuários em cada dia da semana.

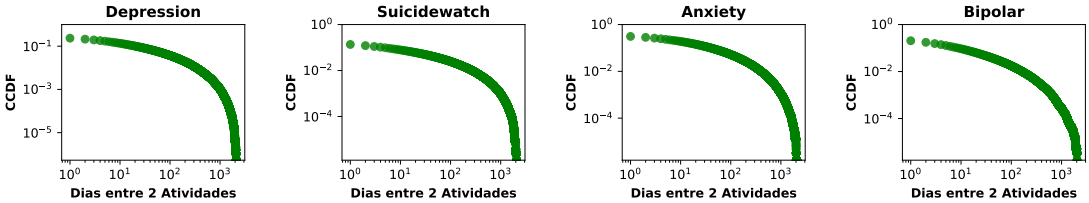



Figura 5. Tempo entre duas atividades (em dias).

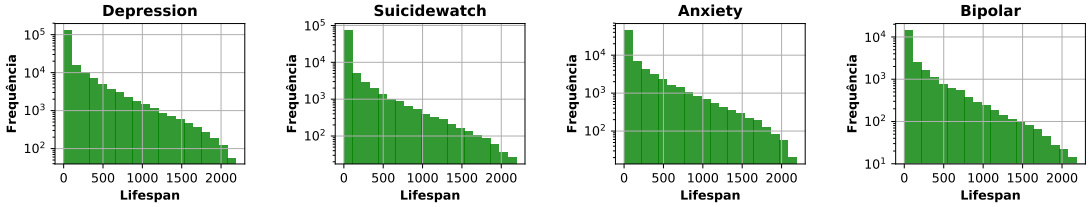



Figura 6. Lifespan de um usuário (em dias).

diferença, em dias, entre duas atividades consecutivas de um mesmo usuário. Observa-se que mais de 80% das atividades consecutivas ocorrem no mesmo dia. Em raríssimas ocasiões, os intervalos chegam a 2000 dias. Esta distribuição decai mais rapidamente que uma lei de potência.

A Figura 6 apresenta o *lifespan* de um usuário em cada comunidade. O *lifespan* foi calculado como a diferença, em dias, entre a primeira e a última atividade do usuário dentro da comunidade. Consequentemente, os usuários que tiveram apenas uma ativi-

dade na comunidade, não foram considerados. Tais usuários representam na comunidade Anxiety 40%, na Bipolar 31,2%, na Depression 42,4% e na SuicideWatch 42,3%. A distribuição do *lifespan* é muito semelhante entre as comunidades. Como os dados analisados correspondem ao intervalo de 2012 a 2017, o *lifespan* mais longo possível é 2190 dias. Observamos que há pouquíssimos usuários com *lifespan* maior que 2000 em cada uma das comunidades, e que a maior parte dos usuários permanece por menos de 100 dias. A mediana do *lifespan* em dias da comunidade Depression é 9,69, SuicideWatch 0,84, Anxiety 25,09 e Bipolar 39,74.

## 4. Metodologia

Nesta seção, apresentamos a metodologia usada para: (i) identificar padrões comuns de atividades entre os usuários das principais comunidades de saúde mental no Reddit e; (ii) compreender como essas comunidades são utilizadas para melhorar o tom emocional de seus usuários (i.e., torná-lo mais positivo). Por fim, definimos os modelos que permitem acompanhar a evolução do tom emocional dos participantes destas comunidades.

### 4.1. Clusterização de usuários segundo perfis de atividade

Identificar grupos de usuários é importante para entender diferentes perfis de atividades encontrados no Reddit. Assim, é possível analisar as características de cada grupo e, consequentemente, como eles respondem às interações (*posts* e comentários) dentro da comunidade. Com esses perfis mapeados, é possível aplicar técnicas de intervenção mais apropriadas para diferentes usuários. Para mapear os usuários nos perfis, definimos os seguintes atributos:

1. **Frequência de Atividades**: Frequência de atividades (*posts* e comentários) por unidade de tempo (dias). Corresponde à razão entre a quantidade de atividades do usuário e o seu *lifespan* (isto é, diferença, em dias, entre a última e a primeira atividade do usuário). Distingue os usuários que são mais ativos em uma comunidade durante os seus *lifespans*.
2. **Iniciativa**: A fração de *threads* das quais o usuário participa que foram disparadas por ele. Assume valor mínimo (igual a 0) quando o usuário participa apenas de threads já existentes, criadas por outros, e valor máximo (igual a 1) quando participa apenas de suas próprias threads. Ou seja, a ideia é verificar se as atividades do usuário se concentram em fazer comentários ou iniciar novas discussões.
3. **Auto-Engajamento**: Fração dos comentários do usuário que foram realizados em árvores de discussão iniciadas por ele. É a razão entre a quantidade de comentários em árvores de discussão iniciadas pelo usuário e o número total de comentários do usuário. Este atributo distingue os usuários que são engajados em suas próprias publicações, ou seja, priorizam comentar seus próprios *posts*. Diferente do atributo "Iniciativa", este atributo considera o número de comentários e não apenas a participação binária (sim ou não) nas threads.

### 4.2. Análise das Interações

Analisamos as interações entre os usuários através do seguinte modelo matemático. Em cada comunidade, as interações foram representadas por um grafo direcionado $G_d =$

$(V, E_d)$, no qual o conjunto de vértices $V$ representa usuários que realizaram um *post* ou um comentário na rede e o conjunto de arestas $E_d$ representa interações entre pares de usuários em $V$. Essas arestas são direcionadas: se o vértice $v$ respondeu a um *post* ou comentário de um vértice $u$, então $v$ aponta para $u$. O grafo modelado possui pesos nas arestas, sendo o peso igual ao número de interações (respostas a *post* ou comentários) que ocorreram entre cada par de vértices em uma determinada direção. Para o cálculo de algumas métricas, utilizamos uma versão não-direcionada $G = (V, E)$ do grafo, em que cada aresta $(u, v) \in E$ indica a existência de $(u, v) \in E_d$ ou $(v, u) \in E_d$. O peso associado a $(u, v) \in E$ é igual à soma dos pesos das arestas entre $u$ e $v$, em ambas as direções, no grafo direcionado.

Para caracterizar o grafo de interações entre os usuários calculamos as seguintes métricas clássicas de redes complexas [Newman 2003]: diâmetro da rede, número de triângulos, *closeness*, excentricidade, grau de entrada e saída e *clustering coefficient*.

### 4.3. Análise do Tom Emocional

Tom emocional é o sentimento geral (positivo e negativo combinado) extraído de um texto. Assim, para mensurar o estado de um usuário, analisamos o texto dos *posts* e comentários coletados utilizando a ferramenta de análise de sentimentos Vader [Hutto and Gilbert 2014]. Após a execução do algoritmo, o texto analisado é classificado, proporcionalmente, nas seguintes categorias léxicas: positiva, negativa ou neutra. Adicionalmente, a ferramenta calcula a composição dos três valores léxicos normalizados entre -1 (extremo negativo) e 1 (extremo positivo), denominado *compound*, a que normalmente se refere por tom emocional.

A análise do tom emocional e a quantificação da sua variação ao longo do tempo são realizadas considerando as árvores de discussão (i.e, *threads*) extraídas das interações entre os usuários dos subreddits considerados. Para analisarmos a variação do tom emocional em cada árvore de discussão isoladamente, consideramos apenas sequências de comentários em que: (i) o usuário inicia a discussão e inclui um comentário antes de se tornar ativo em outra árvore de discussão; (ii) exista pelo menos uma interação com outro usuário e (iii) o intervalo entre duas interações consecutivas seja menor ou igual a 24 horas.

A Figura 7 ilustra a metodologia para seleção das árvores de discussão. O eixo horizontal corresponde ao tempo. O usuário em questão realiza atividades (posts e/ou comentários) nas *threads* 1 a 5 durante o intervalo ilustrado. As atividades realizadas por este usuário são marcadas por marcadores mais grossos, identificados por P no caso de *post*, ou $C_i$ no caso do $i$-ésimo comentário que ele faz na *thread*. Atividades de outros usuários são marcadas por linhas tracejadas. No exemplo, a *thread* 4 não será selecionada para análise porque entre o *post* e seu primeiro comentário na *thread*, ele interagiu na *thread* 2. Por razão análoga, o comentário $C_2$ na *thread* 2 será desconsiderado. Dado que a ordem em que as *threads* são avaliadas impacta no conjunto final de árvores de discussão, estabelecemos como critério avaliar as *threads* em ordem crescente segundo a data de criação do *post*. Após aplicar o critério de seleção apresentado, permanecemos com 87.906 árvores no Depression (27,78% do total), 41.087 no SuicideWatch (37,81% do total), 33.821 no Anxiety (32,64% do total) e 19.267 no Bipolar (33,17% do total).

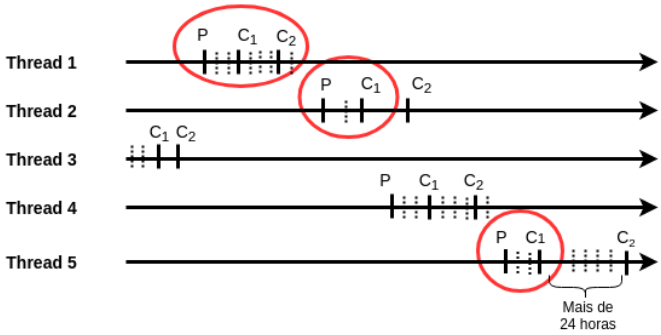



Figura 7. Apenas os trechos das árvores de discussão circulados foram considerados. O usuário $A$ inciou as *threads* 1, 2, 4 e 5. Não consideramos a *thread* 4, pois o usuário ainda realizou uma interação concomitante na *thread* 2.

### 4.4. Modelagem do efeito das interações no tom emocional

Modelamos o efeito das interações em uma thread no tom emocional de um usuário como uma tarefa de predição. Especificamente, consideramos como entradas dos modelos de previsão: o *post* inicial $p$, com tom emocional TE($p$) realizado por um usuário $u$, e o conjunto de comentários da mesma *thread* ($c_1, c_2, ..., c_{last}$), onde o comentário $c_{last}$, com tom emocional TE($c_{last}$), é o último comentário realizado pelo usuário $u$. O resultado do modelo é uma previsão da variação do tom emocional, TE($c_{last}$) - TE($p$), do usuário $u$. A partir deste resultado, será possível verificar se houve alguma mudança do sentimento do usuário após as interações realizadas com os demais usuários da comunidade em questão.

**Representação dos Posts e Comentários.** Para desenvolver a tarefa proposta, é preciso representar posts e comentários através de atributos. Neste trabalho, utilizamos a técnica de *Word Embeddings* (WE), também conhecida como *words vectors* ou *word representations*, que representam palavras como vetores em um espaço latente. Nesse espaço, vetores similares representam palavras com o significado ou função semelhantes.

Para a criação dos *words-embeddings* usamos todos os *posts* e comentários de 2010 a 2017 de cada subreddit. No pré-processamento, utilizamos a ferramenta ekphrasis[11], que realiza a normalização das palavras e correção ortográfica. Aplicamos a normalização no e-mail, telefones e horas. Também removemos as pontuações, caracteres especiais e excesso de espaço. Para remover as *stopwords* utilizamos o *Natural Language Toolkit* (NLTK) [Bird et al. 2009].

Na geração dos *Words Embeddings* (WE) utilizamos a técnica *word2vec* implementada no pacote Gensim do Python.[12] Os parâmetros para execução do modelo foram: arquitetura *SkipGram* [Goldberg 2017]; contexto igual à 5; considerando todas as palavras (que tenham aparecido ao menos uma vez no corpus); e dimensão dos vetores 300. O WE retorna um vetor para cada palavra do *post*/comentário, treinado com o corpus que possuímos. Para agregar os *embeddings* associados a cada publicação, tomamos a média dos *embeddings* das palavras. Portanto, cada publicação é representada por um vetor composto pela média dos *embeddings* de todas as palavras.

[11]https://github.com/cbaziotis/ekphrasis
[12]https://pypi.org/project/gensim/

**Modelos de Previsão.** Neste trabalho, utilizamos o modelo de rede neural Multilayer Perceptron (MLP) [Gardner and Dorling 1998] para prever o tom emocional dos usuários das comunidades analisadas. Escolhemos esse modelo por capturar possíveis relações não-lineares entre o TE e pares de atributos (o que não seria possível com uma regressão linear), sem o elevado número de parâmetros de uma arquitetura mais profunda. Utilizamos 3 camadas ocultas, tendo a primeira 128 neurônios e as outras duas, 256. Foram criadas redes neurais de dois tipos: (i) uma apenas com perceptrons e (ii) outra com camadas de regularização. Para a regularização utilizamos a camada de $Dropout$ com os valores de $0.3$, $0.5$ e $0.7$. Depois dos testes, permanecemos com o valor de $0.5$. Consideramos 5 variações do conjunto de atributos de entrada para cada uma das redes apresentadas, discutidas a seguir. É importante ressaltar que, em todos modelos, desconsideramos o último comentário do autor do *post* na árvore de discussão ao calcular o vetor de entradas para a rede (i.e., as *features*).

- Modelo 1 - Agregação de comentários (dimensão da entrada = 300): a entrada é um vetor de dimensão 300 que é o *embedding* médio de todos os comentários da árvore de discussão.
- Modelo 2 - Agregação de *Post* e Comentários (dimensão da entrada = 300): a entrada é um vetor de dimensão 300 que é a média do *embedding* do *post* com o *embedding* médio de todos comentários da árvore de discussão.
- Modelo 3 - *Post* e agregação de Comentários (dimensão da entrada = 600): a entrada é composta por dois vetores concatenados, sendo um vetor de dimensão 300 que o *embedding* médio de todos comentários da árvore de discussão, e outro vetor de dimensão 300, que é o *embedding* do *post*.
- Modelo 4 - *Post*, agregação de comentários do autor e agregação de outros comentários (dimensão da entrada = 900): a entrada é composta por três vetores, sendo um vetor de dimensão 300 que é o *embedding* médio de todos comentários (excluindo aqueles do usuário que realizou o *post*), outro vetor de dimensão 300, que é o *embedding* médio dos comentários do autor do *post*, e um vetor de dimensão 300, que é o *embedding* do *post*.
- Modelo 5 - Agregação de *post* e comentários do autor e agregação de outros comentários (dimensão da entrada = 600): a entrada é composta por dois vetores, sendo um vetor de dimensão 300, que é o *embedding* médio de todos comentários (excluindo aqueles do usuário que realizou o *post*); e outro vetor de dimensão 300, dado pela média entre o *embedding* do *post* e o *embedding* médio dos comentários do autor do *post*.

A avaliação da acurácia dos diferentes modelos é realizada a partir da métrica de MSE (*Mean Squared Error*).

## 5. Resultados

Nesta seção apresentamos a caracterização dos usuários segundo perfis de atividade, os resultados da análise do tom emocional dos usuários e dos modelos de predição propostos.

### 5.1. Clusterização de usuários segundo perfis de atividade

Agrupamos os usuários com base nos atributos que descrevem suas atividades (frequência de atividades, tipo de atividades e auto engajamento), encontrando quatro clusters em cada

| | Percentual de usuários em cada grupo | | | |
|---|---|---|---|---|
| **Grupo** | **Depression** | **SuicideWatch** | **Anxiety** | **Bipolar** |
| 0 | 17,36% | 28,15% | 22,08% | 17,86% |
| 1 | 0,80% | 3,38% | 0,12% | 0,04% |
| 2 | 13,32% | 8,39% | 11,02% | 14,98% |
| 3 | 68,52% | 60,07% | 66,78% | 67,13% |

Tabela 3. Percentual de usuários em cada grupo por comunidade.

uma das comunidades. O tipo de atividades e o auto engajamento estão associados com a participação dos usuários nas árvores de discussão. Usuários que possuem valores altos de auto engajamento podem ser considerados "egocêntricos", uma vez que sua participação se dá, na maior parte das vezes, em suas próprias árvores de discussão. Por outro lado, usuários que participam de muitas árvores de discussões criadas por usuários podem ser considerados "altruístas". Os usuários com um valor alto no atributo frequências de atividades, são aqueles que são mais participativos na realização de *posts* e comentários, podendo ser considerados "ativos".

Observe que estes atributos são como dimensões independentes: um usuário pode ser bem ativo e ter como foco principal suas próprias *threads*. Contudo, o algoritmo de clusterização irá agrupar os usuários segundo as características que destoam mais intensamente da média.

A Tabela 3 apresenta o percentual de usuários em cada grupo. Como os clusters foram calculados de maneira independente para cada comunidade, reindexamos os grupos encontrados de forma que grupos com mesmo índice representassem usuários com distribuição de atributos similares. Coincidentemente, a proporção entre os grupos ficou semelhante para as diferentes comunidades.

Para a visualização da distribuição de cada atributo, retiramos a padronização. A Figura 8 apresenta a distribuição das características em cada grupo dos subreddits Depression, SuicideWatch, Anxiety e Bipolar.

O grupo 0 contém os usuários com maior auto-engajamento. Assim, o consideramos como o grupo mais egocêntrico. O grupo 1 contém os usuários com maior frequência de atividades por unidade de tempo. Este mesmo grupo possui, exceto pelos *outliers*, um auto-engajamento relativamente baixo, com exceção da comunidade Suicidewatch, na qual 50% dos usuários analisados possuem um nível variável de engajamento em suas árvores de discussão, que estão distribuídos entre o segundo e terceiro quartis do gráfico, sendo que os demais participam somente nas árvores de discussão iniciadas por outros usuários. Além disso, a fração de atividades do tipo *post* é baixo em relação aos demais grupos. Assim, o grupo 1 é predominantemente composto por usuários ativos, que comentam bem mais do que postam e que são altruístas, uma vez que atuam mais comentando outros *posts*. O grupo 2 é marcado por usuários um pouco menos ativos que os dos outros grupos. Além disso, são usuários com o segundo maior auto-engajamento, ou seja, realizam uma grande proporção de comentários em suas próprias publicações. Ademais,

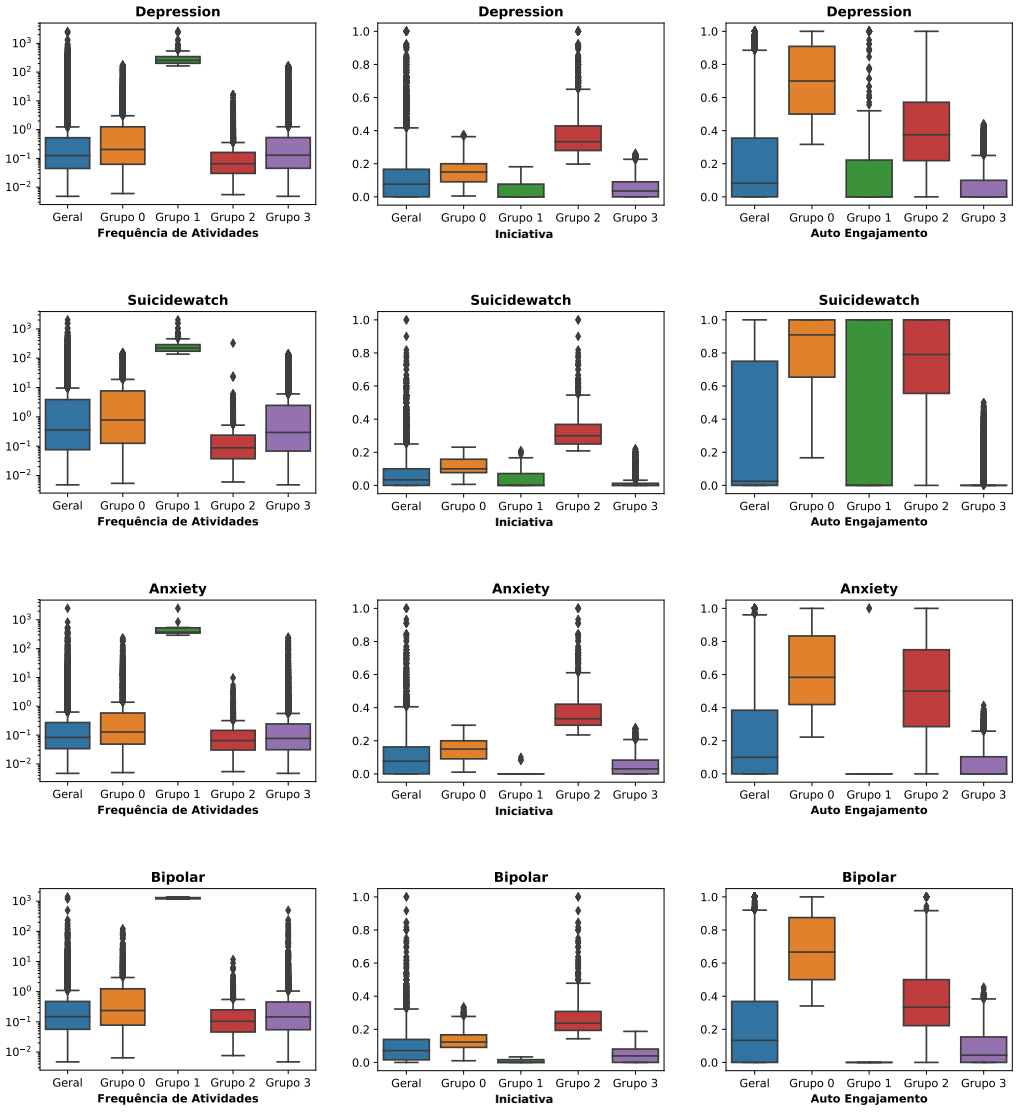



Figura 8. Grupos em cada Subreddit por característica.

é um grupo com a razão mais elevada de *posts*/comentários. Em suma, os usuários menos ativos, relativamente egocêntricos e que mais realizam postagens prevalecem no grupo 2. Os usuários do grupo 3 são os mais altruístas dentre todos os grupos, e se distinguem do grupo 1 por não realizar atividades com frequência acima da média.

Os clusters obtidos para os outros subreddits são semelhantes àqueles obtidos para o subreddit Depression. A única diferença notável é que, nas comunidades Anxiety e Bipolar, o Grupo 1 tem auto-engajamento menor que o Grupo 3. Contudo, estes são os dois grupos menos egocêntricos, independentemente do subreddit analisado.

## 5.2. Análise das Interações

Nosso conjunto de dados é composto por postagens e comentários dos usuários que estavam nas comunidades entre 2012 e 2017. Especificamente nesta análise, focamos

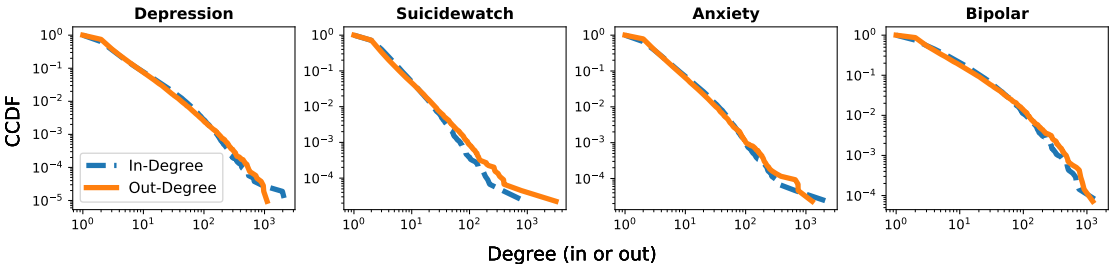



Figura 9. Função de Distribuição Cumulativa Complementar (CCDF) dos graus de entrada e saída do usuário.

nos dados que compreendem o período de Janeiro de 2017 a Dezembro de 2017, uma vez que o objetivo é entender a estrutura mais atual das comunidades deste estudo. Além disso, este ano concentra o maior volume de posts e comentários da base de dados [Silveira Fraga et al. 2018]. No total, obtivemos 261.511 publicações e 1.256.669 comentários de 184.708 usuários únicos. O número total de comentários em cada comunidade é pelo menos 4,8 vezes maior que o número de postagens, o que indica a existência de uma rede de suporte ativa entre os usuários.

A partir de um grafo direcionado $G_d$ com pesos que representam o número de interações do nó A com o nó B, calculamos as distribuições dos graus de entrada e de saída dos vértices. O grau de entrada de um vértice é a quantidade de comentários que o usuário correspondente recebeu em suas publicações. O grau de saída de um vértice é quantidade de publicações feitas por um usuário.

A Figura 9 mostra a Função de Distribuição Cumulativa Complementar (CCDF) dos graus de entrada e saída. Ambas são distribuições de cauda pesada para todos os subreddits. Os maiores valores de grau de entrada e saída são semelhantes, exceto para Depression, em que o maior grau de entrada é cerca de duas vezes maior que o grau de saída. Uma inspeção minuciosa do grau de entrada revela que 63,8% (Depression), 71,5% (Suicide), 66,2% (Anxiety) e 75,2 % (Bipolar) dos usuários receberam pelo menos um comentário em suas postagens. Além disso, observamos que 11,9% (Depression), 8,9% (SuicideWatch), 12,0% (Anxiety) e 20,5% (Bipolar) de usuários comentaram 5 a 20 vezes, sugerindo altos níveis de troca de informações nessas comunidades. Ademais 2,8% (Depression), 1,4% (SuicideWatch), 2,0% (Anxiety) e 8,8% (Bipolar) de usuários contribuem com mais de 20 comentários.

Ambos os graus de entrada e saída são medidas de diversidade de interação. No entanto, também estamos interessados em sua intensidade. Nesse caso, analisamos os pesos das arestas do grafo, que representam o número de comentários feitos/recebidos por um usuário. Cerca de 84,7% dos membros interagem um com o outro apenas uma vez no subreddit Depression, 73,6% no SuicideWatch, 85,7% no Anxiety e 81,5% no Bipolar. Um número não desprezível de pares interage de 2 a 20 vezes: 15,2% no Depression, 26,3% no SuicideWatch, 14,3% no Anxiety e 18,5% no Bipolar.

Uma outra maneira de quantificar o volume de *posts* e comentários em um subreddit é mensurar a quantidade de interações entre pares de usuários. Dizemos que houve

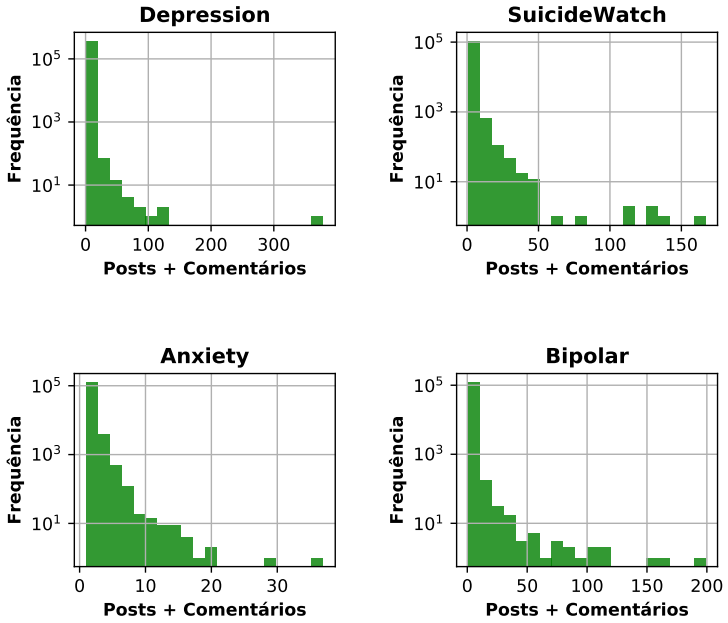



Figura 10. Histograma do número de arestas quantizado por volume de interação.

uma interação entre um par de usuários $(i, j)$ quando $i$ comenta em uma publicação de $j$ ou vice-versa. A Figura 10 apresenta o histograma contendo o número de pares encontrados quantizados por números de interações. Podemos observar que a grande maioria dos usuários estabelece poucos diálogos entre si (primeira coluna dos histogramas apresentados). Este resultado corrobora a filosofia da rede social Reddit, cujo o objetivo principal é o engajamento em torno de conteúdos em que os usuários possuem interesse, independentemente dos usuários que participam da discussão (dado que a identidade dos usuários sequer é revelada, em muitos casos).

Calculamos várias métricas estruturais a partir do grafo de interação de cada subreddit. Mais especificamente, calculamos o diâmetro da rede (maior distância entre nós), transitividade (fração de todos os triângulos possíveis que existem em $G$), além de estatísticas – média, mediana, coeficiente de variação (CV) – das métricas dos nós [Newman 2003], especificamente, excentricidade, *closeness*, *clustering coefficient* e número de triângulos. A Tabela 4 mostra os resultados obtidos para cada métrica. No geral, observamos que os padrões de interação são muito semelhantes entre os subreddits. A baixa transitividade combinada com alto diâmetro e excentricidade do nó corrobora o fato de que as interações do usuário são baseadas principalmente no conteúdo das postagens e comentários, independentemente dos usuários que os geram. Este comportamento é encorajado pelo Reddit, que não tem nenhum filtro de conteúdo baseado em amizades e proíbe os "anéis de votação"em seu termo de serviços. Nós conjecturamos que esta é também a razão para os baixos valores de *closeness*. Também observamos uma grande variação (em termos de CV) na distribuição do número de triângulos, indicando que esse número varia substancialmente em diferentes nós. Essa é uma consequência direta da grande variação no número de *posts* e comentários realizados por diferentes usuários.

| Subreddit | Métrica Grafo | | Métrica por nó | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Excentricidade | | | Closeness | | | Clustering Coeff. | | | # Triângulos | | |
| | Diâm. | Trans. | $X_{50}$ | $\bar{X}$ | CV | $X_{50}$ | $\bar{X}$ | CV | $X_{50}$ | $\bar{X}$ | CV | $X_{50}$ | $\bar{X}$ | CV |
| Depression | 12 | 7e-3 | 8 | 8,47 | 0,07 | 0,13 | 0,09 | 0,84 | 0 | 0,03 | 4,50 | 0 | 2,30 | 18,33 |
| SuicideWatch | 15 | 1e-3 | 10 | 10,17 | 0,08 | 0,10 | 0,08 | 0,74 | 0 | 0,02 | 5,51 | 0 | 0,32 | 19,11 |
| Anxiety | 12 | 4e-3 | 8 | 8,27 | 0,08 | 0,13 | 0,09 | 0,80 | 0 | 0,03 | 4,30 | 0 | 0,89 | 16,23 |
| Bipolar | 8 | 4e-2 | 6 | 5,65 | 0,11 | 0,23 | 0,18 | 0,60 | 0 | 0,13 | 1,73 | 0 | 5,80 | 8,34 |

Tabela 4. Métricas do grafo de interação do usuário por subreddit ($X_{50}$: mediana, $\bar{X}$: média, CV: coeficiente de variação).

Em suma, nossa análise mostra que nesses quatro subreddits relacionados à discussão sobre saúde mental, o modelo de interação é centrado em torno do conteúdo das postagens e comentários, e não dos usuários. Isto é positivo, pois este modelo ajuda os novos usuários a iniciar sua participação na rede, uma vez que impede a formação de grupos muito unidos.

### 5.3. Análise do Tom Emocional

Recapitulando a notação introduzida previamente, seja TE($p$) o tom emocional de um *post* p; TE($c_{last}$) o tom emocional do último comentário feito pelo autor de uma *thread* naquela mesma *thread*; e TE($c_{all}$) comentários dos outros usuários, excluindo o autor do *post*.

**Características TE Árvores de Discussão.** A Figura 11 apresenta a distribuição de, TE($p$), TE($c_{last}$) e $\overline{\text{TE}(c_{all})}$, i.e, TE do *post*, do último comentário do autor do *post* e a média de TE dos comentários da árvore de discussão, sem incluir $c_{last}$. O gráfico apresentado é do tipo violino. Apesar da semelhança com o gráfico do tipo boxplot, o gráfico do tipo violino permite a análise da densidade dos valores de tom emocional: quanto mais largo o formato, maior é a concentração de pontos naquela região. O ponto branco indica a mediana. A barra mais grossa é o intervalo interquartil.

É possível observar que, em todos subreddits, o 1º quartil e a mediana referentes ao *post* estão bem abaixo das respectivas medianas dos últimos comentários do autor e dos comentários dos outros usuários. Além disso, o 3º quartil referente ao *post* também se encontra abaixo daquele referente à média dos comentários, exceto no subreddit Bipolar. Em geral, os comentários tendem a ser mais positivos que os *posts*, exceto para o subreddit Bipolar, que apresenta também um grande volume de *posts* com o tom emocional mais positivo. Note que, em geral, há maior concentração de *posts* na região de TE negativo. Por outro lado, no caso do último comentário feito pelo autor da *thread*, há maior concentração na região de TE positivo.

Além disso, a mediana do TE médio dos comentários (i.e., de $\overline{\text{TE}(c_{all})}$) é menor que o TE do último comentário do autor do *post*, exceto no subreddit SuicideWatch. Esta análise sugere que as interações do autor de um *post* com outros usuários, ao receber comentários com TE positivo, podem contribuir para o aumento do seu TE.

**Relação entre TE do Post e dos Comentários.** Para entender como os usuários que participam de uma árvore de discussão interagem com o autor do *post*, apresentamos a

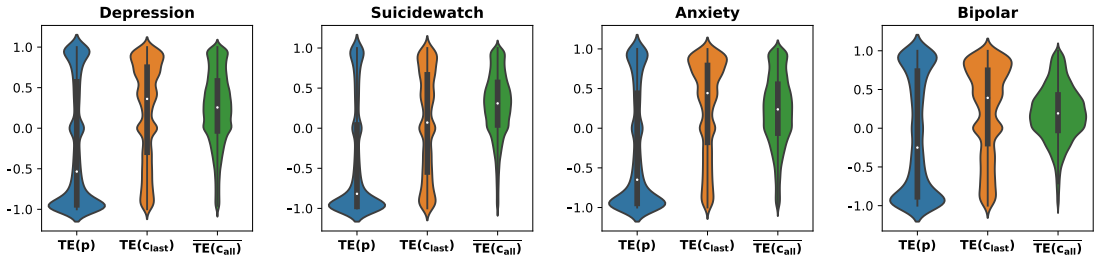



Figura 11. Distribuição dos valores do Tom Emocional dos posts e comentários.

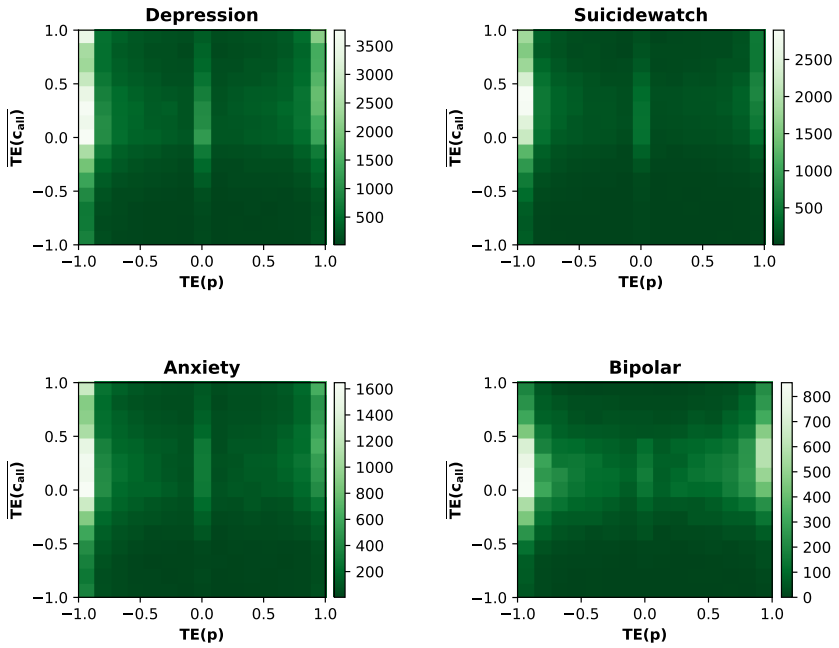



Figura 12. Tom Emocional dos Posts x Tom Emocional Médio dos Comentários.

Figura 12, que mostra um mapa de calor do número de árvores de discussão, agrupadas de acordo com o TE do *post* (eixo x) e com o TE médio dos comentários da árvore de discussão, excluindo-se o *post* inicial. É possível observar que quando o *post* tem TE extremamente negativo (-1), a média dos comentários dos outros usuários na árvore de discussão tende a ser mais positiva (valores acima 0).

**Relação entre TE do Post e do Último Comentário.** Averiguamos também se ocorre mudança do TE entre o momento em que um usuário escreve um *post* e o momento em que faz seu último comentário na árvore de discussão, utilizando um mapa de calor semelhante ao anterior, mostrado na Figura 13. Note que, autores de *posts* extremamente negativos (-1) tendem a escrever comentários mais positivos ao final da *thread*. Esta variação sugere que os comentários realizados por outros usuários podem ajudar os usuários que se encontram em situações difíceis.

**Relação entre TE do Último Comentário e dos demais Comentários.** O passo seguinte foi investigar se o TE do último comentário do autor tem relação com a média do TE dos

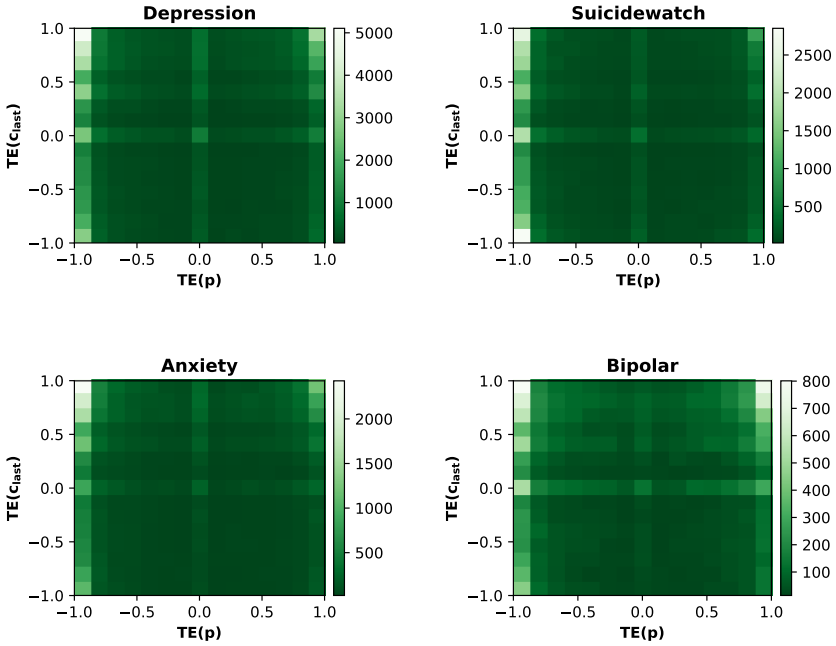



Figura 13. Tom Emocional dos Posts x Tom Emocional do Último Comentário.

comentários da árvore de discussão (excluindo-se o *post* inicial), calculando-se o mapa de calor mostrado na Figura 14. Note que, em todos subreddits, um TE positivo no último comentário do autor do *post* está relacionado a um TE médio positivo entre os comentários *thread*. No SuicideWatch também ocorre uma grande quantidade de últimos comentários negativos por parte dos autores da *thread*, mesmo quando estão relacionados a uma árvore de discussão com TE positivo. Isto pode ser um indício de que existem alguns usuários nesta comunidade que dificilmente se beneficiam de uma melhoria do tom emocional, ou seja, de um aumento da positividade nos textos das suas atividades.

**Relação entre TE dos Comentários e da Diferença do TE do último comentário e o Post.** Como observado anteriormente, o TE dos comentários das árvores de discussões estão positivamente correlacionados com o TE do último comentário. Consequentemente, podem estar relacionados com uma variação positiva, considerando-se a diferença entre o TE do *post* e o TE do último comentário. Para investigar melhor estas relações, analisamos a diferença, para cada árvore de discussão, entre o tom emocional do último comentário do autor e do *post*. A Figura 15 apresenta a distribuição desta diferença. Observe que ocorre maior concentração nos valores acima de zero, mostrando que o usuário, na maioria das vezes, "melhora" ao final da *thread*, em relação ao TE. Analisamos então a relação entre a média do TE dos comentários e a variação do TE observada pelo autor do *post*. Na Figura 16, notamos que quando a média é positiva, existe uma diferença positiva entre o TE do último comentário e o *post*, que pode ser observada no canto superior direito. Note que quando a média é negativa, não existe um padrão claro para a diferença do TE.

Em resumo, as análises realizadas fornecem um conjunto de observações relevan-

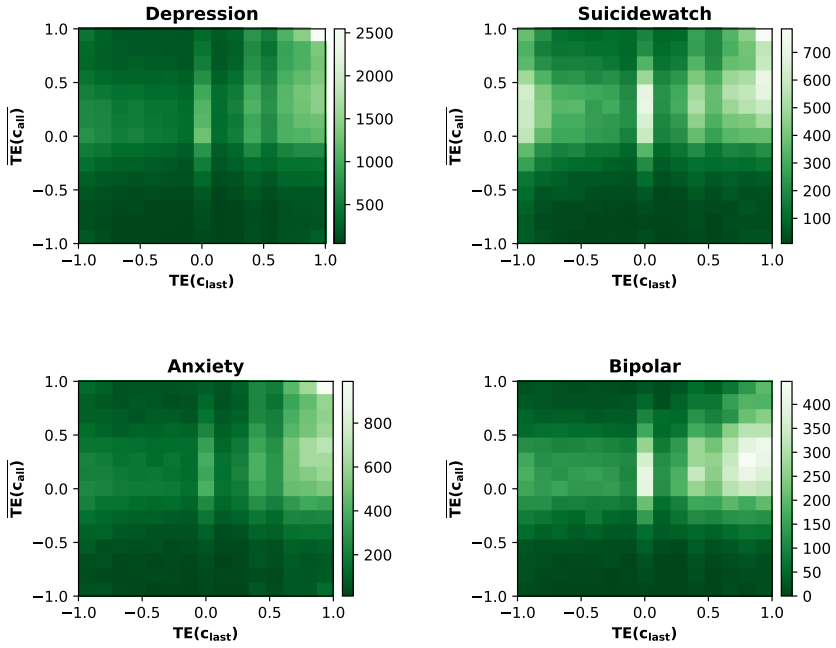



Figura 14. Média do Tom Emocional dos Comentários x Tom Emocional do último comentário.

tes a respeito da dinâmica do tom emocional de um usuário ao participar em uma árvore de discussão. Investigamos vários aspectos do TE dentro da árvore de discussão: nos *posts*, nos comentários, no último comentário do autor do *post* e na diferença do TE. Um dos resultados mais importantes é a existência de correlação positiva entre o TE do último comentário do autor da thread e o TE dos comentários dos demais usuários na árvore de discussão. Além disso, observamos que um comentário influencia o TE do próximo comentário, o que é um indício de que o aspecto temporal pode impactar na evolução do comportamento do usuário dentro da comunidade. A seguir, avaliamos a acurácia dos modelos projetados para prever a diferença do TE do usuário ao longo tempo, dentro da árvore de discussão.

### 5.4. Modelagem do efeito das interações no tom emocional

Nesta seção, apresentamos os resultados da acurácia dos modelos de previsão propostos.

**Métrica de avaliação.** Seja $D$ a variável aleatória que representa a variação do TE sofrida pelo autor de uma árvore de discussão escolhida uniformemente ao acaso, definida por $D = \text{TE}(c_{last}) - \text{TE}(p)$. Sejam ainda $D_i$ a variação do TE associada à *thread* $i$ e $\hat{D}_i$ a respectiva previsão, retornada por um dado modelo. Considerando o *Mean Squared Error* (MSE) como medida de erro, podemos avaliar os modelos utilizando a equação

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(D_i - \hat{D}_i)^2. \tag{1}$$

As variáveis $D_1, \ldots, D_n$ representam a variação de TE associada às *threads* do conjunto de teste. Para calcular as métricas de erro, utilizamos *cross-validation* com 5 *folds*.

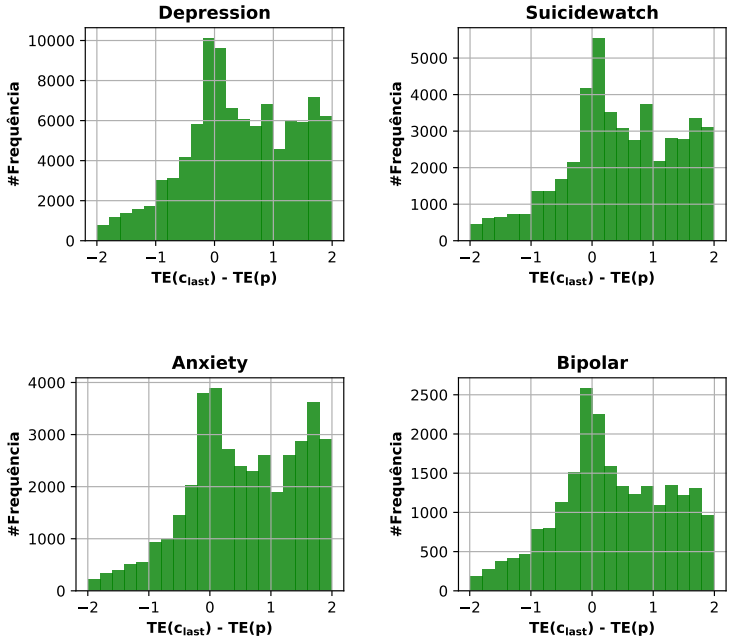



Figura 15. Histograma da Diferença do Tom Emocional entre comentário e post

**Baselines.** Definimos três baselines para avaliar a eficácia dos modelos propostos:

- Baseline 1: O primeiro baseline é o modelo que sempre retorna 0 para a variação do TE, ou seja, que assume que o usuário manteve seu TE na árvore de discussão. Neste caso, o MSE do Baseline 1 é dado por: $\text{MSE}(B_1) = E[D^2]$.
- Baseline 2: O segundo baseline é o modelo que sempre retorna a variação média do TE, $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$. Neste caso, o erro de previsão é calculado por: $\text{MSE}(B_2) = E[(D - \bar{D})^2]$, ou seja, é a variância de $D$.
- Baseline 3: O terceiro baseline utilizado foi uma regressão linear considerando como atributos o total de comentários e a média do TE da árvore de discussão (excluindo-se àqueles de quem fez o *post*) para prever a diferença do TE. O erro correspondente é calculado através da Eq. 1.

A Tabela 5 apresenta o MSE obtido para cada baseline avaliado em cada comunidade. Observe que os MSE's do baseline 3 são menores que os MSE's dos baselines 1 e 2, mas que a diferença em relação ao modelo 2 é pequena, indicando a dificuldade de prever $D$ a partir de $\overline{\text{TE}(c_{all})}$. Assim, nossos modelos terão o objetivo apresentar resultados mais precisos que os resultados do baseline 3.

**Resultados.** Testamos cada modelo proposto na Seção 4.4 com e sem regularização na comunidade Bipolar, a fim de selecionar o modelo a ser usado em todos os subreddits. A Tabela 6 apresenta os resultados das 10 configurações (5 modelos $\times$ uso/não-uso de regularização) para a comunidade Bipolar. Observe que os resultados dos modelos sem regularização são ligeiramente piores. Em particular, o modelo 1 sem regularização resulta em erro maior que o baseline 3. Isso acontece porque a regularização evita o *overfitting*. Assim escolhemos, dentre os modelos com regularização, o modelo 4, por resultar

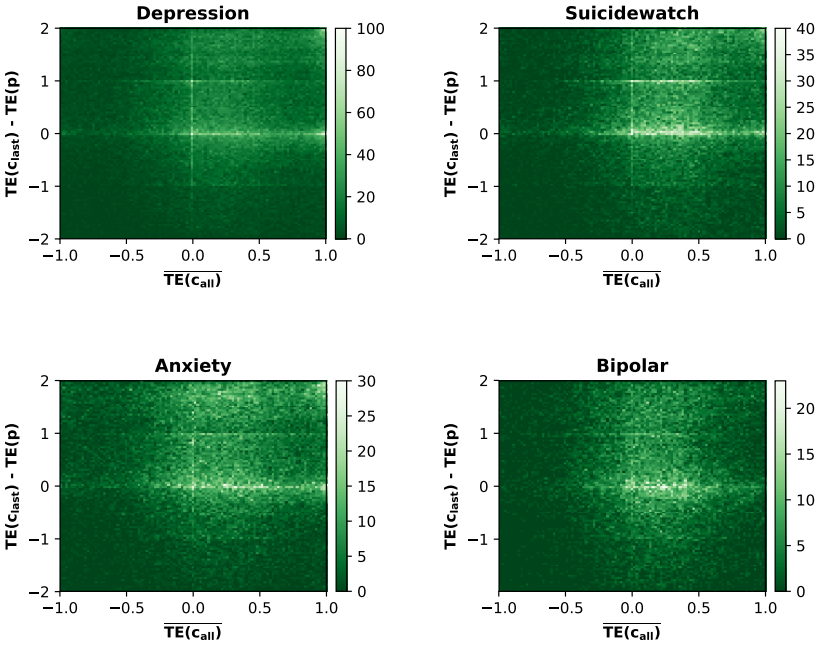



Figura 16. Média do Tom Emocional dos Comentários x Diferença do tom emocional entre último comentário e post.

| | Mean Squared Error (MSE) | | |
|---|---|---|---|
| **Subreddit** | **Baseline 1** | **Baseline 2** | **Baseline 3** |
| Depression | 1,023 | 0,854 | 0,848 |
| SuicideWatch | 1,029 | 0,843 | 0,839 |
| Anxiety | 1,062 | 0,818 | 0,813 |
| Bipolar | 0,964 | 0,819 | 0,811 |

Tabela 5. MSE's dos Baselines para cada comunidade.

no menor MSE dentre os demais, sendo este inclusive bem menor que o do baseline 3. Uma possível explicação para o melhor desempenho do modelo 4 na tarefa de previsão proposta, está na separação entre o *post*, os comentários do autor da *thread* e os comentários dos outros usuários, isolando diferentes tipos de tons emocionais inerentes às atividades consideradas: por exemplo, o *post* tem o tom mais negativo e normalmente é um pedido de ajuda, os comentários dos outros usuários costumam ter o tom mais positivo e tem o papel de oferecer suporte e conselhos, já os comentários do autor do *post* podem refletir um processo de melhora. Isto indica que separar os textos em diferentes vetores de entrada é importante para a previsão da diferença do TE.

Aplicamos o modelo 4 às demais comunidades, Depression, SuicideWatch e Anxiety. A Tabela 7 apresenta os resultados. A maior diferença percentual no MSE entre comunidades é de apenas 1,02% (entre as comunidades Bipolar e Depression), mostrando que os resultados para as quatro comunidades são similares. Em comparação com o Baseline 3, o modelo 4 obteve melhores resultados de acurácia para todas as comunidades: na

| | MSE | | | |
|---|---|---|---|---|
| | SEM regul. | | COM regul. | |
| | Média | Desvio | Média | Desvio |
| **Modelo 1** | 0,8186 | 0,0236 | 0,8061 | 0,0192 |
| **Modelo 2** | 0,784 | 0,0182 | 0,7844 | 0,0186 |
| **Modelo 3** | 0,671 | 0,0084 | 0,6961 | 0,0446 |
| **Modelo 4** | 0,6781 | 0,0205 | 0,6651 | 0,0191 |
| **Modelo 5** | 0,7174 | 0,0121 | 0,7295 | 0,0203 |

Tabela 6. Resultados MSE das variações dos Modelos MLP's para a comunidade Bipolar. Baseado em 5 folds em termos de média e desvio padrão.

| | MSE com regul. | |
|---|---|---|
| **Subreddit** | **Média** | **Desvio** |
| Depression | 0,6719 | 0,0225 |
| SuicideWatch | 0,6707 | 0,0257 |
| Anxiety | 0,6665 | 0,0264 |
| Bipolar | 0,6651 | 0,0191 |

Tabela 7. Resultados do MSE Modelo MLP. Baseado em 5 *folds* em termos de média e desvio padrão.

comunidade Depression, a diminuição do MSE é de 20,79%, no SuicideWatch 20,04%, no Anxiety 18,02% e no Bipolar 18,02%.

Em suma, os modelos propostos neste trabalho são eficazes para capturar a variação do TE do usuário em uma árvore de discussão. Esses modelos podem ser utilizados no monitoramento de conversas, por exemplo, em redes sociais de suporte a usuários com problemas de saúde mental. Ao identificar que o usuário não está melhorando seu tom emocional, especialistas podem intervir para evitar situações extremas.

## 6. Conclusão

Dados da Organização Mundial da Saúde alertam para o aumento do total de pessoas no mundo que sofrem de algum tipo de transtorno de saúde mental. A combinação entre recursos escassos (principalmente entre países de economia mais instável), o estigma social associado aos transtornos mentais e a resistência em pedir ajuda, faz com que muitas pessoas que sofrem desses transtornos não sejam ajudadas da melhor maneira possível, levando a situações drásticas e irremediáveis, como o suicídio. Assim, diferentes ferramentas para auxiliar pessoas que passam por problemas relacionados à saúde mental estão sendo exploradas nos últimos anos. Em particular, podemos ressaltar o papel das redes sociais online. Neste contexto, este artigo apresenta uma análise detalhada do tom emocional de participantes de comunidades relacionadas à saúde mental no Reddit. A partir dos resultados desta análise, apresentamos diferentes modelos preditivos que capturam a variação do tom emocional destes usuários.

A caracterização realizada mostra que a maior parte das interações entre os usuários são realizadas na segunda-feira, possivelmente influenciados pela rotina semanal, que gera sentimentos negativos para várias pessoas. Encontramos também grupos de usuários que desempenham papéis semelhantes dentro do Reddit. Esses grupos foram definidos pelo nível de interação dos usuários nas suas próprias árvores de discussão e nas de outros usuários. Percebemos que existem usuários que interagem apenas em suas próprias publicações, possivelmente procurando ajuda/conselho. Há outros que não iniciam discussões (*posts*). No entanto, estes usuários comentam, com grande frequência, os *posts* de outros usuários, o que os caracteriza como usuários possivelmente conselheiros, que tentam ajudar os usuários que passam por alguma situação particular.

As análises apresentadas neste trabalho mostram que as interações entre os usuários de RSOs relacionadas aos transtornos de saúde mental auxiliam na melhoria das suas condições de saúde, uma vez que o TE dos usuários tendem a sofrer variações positivas ao longo das interações realizadas. Os nossos modelos preveem com boa acurácia a variação do TE dos usuários participantes destas comunidades.

Como os diversos trabalhos que utilizam dados de redes sociais online, nosso trabalho possui como principal limitação a impossibilidade de verificar se uma variação positiva no TE está correlacionada com uma melhora real no estado emocional do usuário, já que não temos um contato direto com o usuário para averiguação. No entanto, nossos modelos podem ser aplicados para monitorar a evolução do tom emocional de tais usuários durante as discussões realizadas nestas comunidades: caso o TE não esteja evoluindo positivamente, intervenções podem ser realizadas por profissionais da área, evitando situações extremas (por exemplo, o suicídio).

**Agradecimentos.** Este trabalho foi realizado com apoio financeiro do CNPq, FAPEMIG, CAPES.

## Referências

[Baba et al. 2019] Baba, T., Baba, K., and Ikeda, D. (2019). Detecting mental health illness using short comments. In *International Conference on Advanced Information Networking and Applications*, pages 265–271. Springer.

[Ballona et al. 2015] Ballona, P. R., Prado, P., de Almeida, J. M., and Marques-Neto, H. T. (2015). Analyzing the influence of pope's tweets on his followers' mood. In *Proceedings of the 21st Brazilian Symposium on Multimedia and the Web*, WebMedia '15, page 93–100.

[Barney et al. 2006] Barney, L. J., Griffiths, K. M., Jorm, A. F., and Christensen, H. (2006). Stigma about depression and its impact on help-seeking intentions. *Australian & New Zealand Journal of Psychiatry*, 40(1):51–54.

[Bird et al. 2009] Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

[Blair and Abdullah 2018] Blair, J. and Abdullah, S. (2018). Supporting constructive mental health discourse in social media. In *PervasiveHealth '18*.

[Cranford et al. 2006] Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., and Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, 32(7):917–929.

[Cunha et al. 2017] Cunha, T., Weber, I., and Pappa, G. (2017). A warm welcome matters!: The link between social feedback and weight loss in/r/loseit. In *WWW'17 Companion*, pages 1063–1072. International World Wide Web Conferences Steering Committee.

[De Choudhury 2013] De Choudhury, M. (2013). Role of social media in tackling challenges in mental health. In *Proceedings of the 2nd international workshop on Socially-aware multimedia*, pages 49–52. ACM.

[De Choudhury and De 2014] De Choudhury, M. and De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the International AAAI Conference on Web and Social Media*. ICWSM.

[Gardner and Dorling 1998] Gardner, M. W. and Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636.

[Gkotsis et al. 2016] Gkotsis, G., Oellrich, A., Hubbard, T., Dobson, R., Liakata, M., Velupillai, S., and Dutta, R. (2016). The language of mental health problems in social media. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*.

[Goldberg 2017] Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

[Gruda and Hasan 2019] Gruda, D. and Hasan, S. (2019). Feeling anxious? perceiving anxiety in tweets using machine learning. *Computers in Human Behavior*, 98:245–255.

[Hutto and Gilbert 2014] Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

[Islam et al. 2018] Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., and Ulhaq, A. (2018). Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6(1):8.

[Newman 2003] Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

[Pellert et al. 2020] Pellert, M., Schweighofer, S., and Garcia, D. (2020). The individual dynamics of affective expression on social media. *EPJ Data Science*, 9(1).

[Sahota and Sankar 2019] Sahota, P. K. and Sankar, P. L. (2019). Bipolar disorder, genetic risk, and reproductive decision-making: A qualitative study of social media discussion boards. *Qualitative health research*.

[Shen and Rudzicz 2017] Shen, J. H. and Rudzicz, F. (2017). Detecting anxiety through reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 58–65.

[Silveira et al. 2020] Silveira, B., da Silva, A. P., and Murai, F. (2020). Modelos de previsão do tom emocional de usuários em comunidades de saúde mental no reddit. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 13–24, Porto Alegre, RS, Brasil. SBC.

[Silveira et al. 2018] Silveira, B., da Silva, A. P. C., and Murai, F. (2018). Análise de comunidades de suporte a transtornos de saúde mental do reddit. In *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*. SBC.

[Silveira Fraga et al. 2018] Silveira Fraga, B., Couto da Silva, A. P., and Murai, F. (2018). Online social networks in health care: A study of mental disorders on reddit. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 568–573.

[Stone et al. 1985] Stone, A. A., Hedges, S. M., Neale, J. M., and Satin, M. S. (1985). Prospective and cross-sectional mood reports offer no evidence of a"blue monday"phenomenon. *Journal of Personality and Social Psychology*, 49(1):129.

[Weerasinghe et al. 2019] Weerasinghe, J., Morales, K., and Greenstadt, R. (2019). "because... i was told... so much": Linguistic indicators of mental health status on twitter. *Proceedings on Privacy Enhancing Technologies*, 2019(4):152–171.

[Wolohan et al. 2018] Wolohan, J., Hiraga, M., Mukherjee, A., Sayyed, Z. A., and Millard, M. (2018). Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 11–21, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[Wongkoblap et al. 2019] Wongkoblap, A., Vadillo, M. A., and Curcin, V. (2019). Modeling depression symptoms from social network data through multiple instance learning. *AMIA Summits on Translational Science Proceedings*, 2019:44.