Análise de sentimentos: identificando sentimentos em comentários da Rede Humaniza SUS

Eduardo Alves Silva[a] *e Luis Felipe Rosa de Oliveira*[b]

Resumo: A análise de sentimentos e áreas correlatas ao processamento de linguagem natural tem-se tornado cada vez mais comuns em diversos contextos, seja no empresarial ou acadêmico. Esse tipo de análise facilita a compreensão a respeito de opiniões e sentimentos de diferentes indivíduos em relação a determinado produto ou temática. Ao levar em consideração a crescente difusão das redes sociais e a interação entre usuários nessas redes, a captação de dados para análise de sentimento se tornou mais simples do que era anos atrás, onde frequentemente eram utilizados questionários ou formulários para captar a opinião de uma pessoa a respeito de algo. Com esse intuito e com a alta densidade de informação da rede humaniza SUS (RHS), a primeira rede social brasileira criada a partir de uma política pública, voltada para o tema da humanização da saúde. A rede humaniza SUS torna-se um ambiente de estudo promissor para análise de sentimentos, fazendo-se uso de uma abordagem não supervisionada que visa o uso de um dicionário ou léxico de sentimentos criado com palavras em Português, e com a possibilidade de aplicação utilizando a linguagem de programação python, para apresentar a classificação de sentimentos dos comentários feitos pelos usuários da rede.

Palavras-chave: Rede Humaniza SUS. Análise de Sentimentos. Léxico de Sentimentos. Processamento de Linguagem Natural.

Sentiment Analysis: identifying sentiment in comments on Humaniza SUS network

Abstract: The sentiment analysis and areas related to the natural language processing has become increasingly common in several contexts, be it business or academic. This type of analysis facilitates the understanding of the opinions and feelings of different individuals regarding a given product or thematic. By taking into account the growing diffusion of social networks and the interaction among users in these networks, capturing data for feeling analysis has become simpler than it was years ago, where questionnaires or forms were often used to capture a person's opinion about something. With this intention and with the high density of information of the humaniza SUS network (RHS), the first Brazilian social network created from a public policy, focused on the humanization of health. The humaniza SUS network becomes a promising study environment for analysis of feelings, making use of an unsupervised approach that aims to use a dictionary or sentiment lexicon created with words in Portuguese, and with the possibility of application using the programming language python, to present the classification of feelings of the comments made by the users of the network.

Keyword: Humaniza SUS Network. Sentiment Analysis. Sentiment Lexicon. Natural Language Processing.

a Universidade Nova de Lisboa (NOVAIMS). E-mail: easilva91@gmail.com
b Universidade Federal de Goías (UFGO). E-mail: luisprf@gmail.com. Currículo: http://lattes.cnpq.br/6498992926514286

1 Introdução

A crescente popularização das redes sociais, blogs e plataformas que incentivam a participação e colaboração de usuários, tornou-se comum permitindo o compartilhamento de mensagens que refletem suas opiniões e por vezes sentimentos a respeito de determinada temática ou produto.

Nesse sentido, a análise de sentimentos, ou mineração de opinião, corresponde ao problema de tentar identificar ou extrair emoções, opiniões ou pontos de vista expressados em um texto (DUARTE, 2013).

Liu (2012) menciona que esse tipo de análise está ligado a linguística e ao processamento de linguagem natural (PLN), gerando grande impacto nessa área do conhecimento, podendo ter grandes reflexos também em estudos sobre ciências políticas, economia, e ciências sociais, uma vez que são afetados pela opinião das pessoas.

O uso da análise de sentimentos é extenso, sendo usado para verificar o quanto um produto é bem aceito, de acordo com as opiniões de usuários, ou para compreender o sentimento gerado por uma publicação no Facebook, ou uma hashtag no Twitter. Essa variabilidade torna o uso da análise de sentimentos e opinião, um amplo campo de estudo.

Para a definição de sentimentos em um texto, normalmente é utilizado um léxico de sentimento, servindo como um dicionário de polaridade (FREITAS; VIEIRA, 2015), que agrega palavras que comumente expressam sentimentos positivos ou negativos (LIU, 2012). Atualmente os léxicos e modelos de análise de sentimento na língua inglesa são bastante comuns e assertivos em relação a análise de opinião e sentimento, no entanto, em Português são conhecidos quatro léxicos: OpLexicon, SentiLex, Brazilian Portuguese Linguistic Inquiry and Word Count (LIWC) e Onto.PT (FREITAS; VIEIRA, 2015).

O presente trabalho, tem por objetivo identificar os sentimentos expressados por comentários, fazendo uso do OpLexicon, da primeira rede social do Brasil criada no âmbito de uma política pública. Trata-se da rede humaniza SUS (RHS)1 que se encontra em atividade desde o ano de 2008, atualmente com mais de 30 mil usuários por todo o Brasil, onde muitos são responsáveis pelas mais de 14 mil publicações existentes na rede e os quase 40 mil comentários, existentes nessas publicações.

A rede humaniza SUS, traz consigo uma temática bastante específica, a humanização da saúde. Nesse sentido, foi verificado que tipo de sentimentos/opiniões os usuários da rede expressam a partir dos comentários, servindo também como um experimento para o uso de análise de sentimentos em Português.

2 Análise de Sentimentos

A análise de sentimento é, talvez, a aplicação mais popular da análise de texto, com um grande número de tutoriais, sites e aplicativos que se concentram em analisar o sentimento de vários recursos textuais, desde pesquisas corporativas até análises de crítica de filmes (SARKAR, 2016).

Em sua utilização a análise de sentimento tem uma série de fatores a serem considerados para que se consiga resultados que possam ser considerados coerentes, passando pela fase de pré-

processamento do texto, até o uso de abordagens para identificação de sentimentos, potencialmente utilizando métodos supervisionados que envolvem aprendizado de máquina e não supervisionados, que faz uso de bancos de dados, ontologias ou léxicos de sentimento.

Um dos principais objetivos da análise de sentimentos é analisar um determinado gênero de texto para compreender o que ele expressa. Nesse sentido alguns fatores devem ser considerados, como a polaridade do sentimento e sua subjetividade, de acordo com Sarkar (2016) a análise de sentimentos funciona melhor em textos que têm um contexto subjetivo (opiniões) do quem em textos objetivos (fatos).

Vale ressaltar que, a simples busca por palavras como "bom" ou "ruim" não é o suficiente para expressar o sentimento em um texto (DOSCIATTI; FERREIRA, 2013).

2.1 Léxico de Sentimento

Um léxico de sentimento pode ser um dicionário, vocabulário ou conjunto de palavras que tem uma polaridade positiva ou negativa atribuída (SARKAR, 2016). Para o presente trabalho foi utilizado um método não supervisionado. Sendo assim utilizado o léxico OpLexicon.

O OpLexicon é constituído de um total de 32.191 itens (24.475 adjetivos e 6.889 verbos, Tabela 1), tendo sua construção feita com base em textos jornalísticos e resenhas de filmes escritas em Português do Brasil, além do uso de tesauros e a tradução do léxico de opinião em inglês (SOUZA et al, 2012).

Tabela 1 – Quantidade de palavras por tipo OpLexicon.

Tipo	**Quantidade**
adjetivo (adj)	24.475
verbo (vb)	6.889
hashtag	471
vb (verbo) det (determinante) n (nome) prp (preposição)	103
vb (verbo) n (nome) prp (preposição)	91
vb (verbo) adj (adjetivo)	74
emoticons	66
vb (verbo) adv (advérbio)	22

Fonte: OpLexicon v3.0.

3 Procedimentos Metodológicos

O primeiro passo para se atingir o objetivo proposto foi captar os comentários da RHS em um recorte de 9 anos, a partir do banco de dados foram extraídos comentários de 2008 até o fim de 2017, o que gerou um total de 36.346 comentários.

Após essa etapa, foi iniciado o tratamento do texto dos comentários, utilizando a linguagem de programação python, juntamente com as etapas de sumarização e normalização de texto.

Segundo Sakar (2016) a normalização de texto é um processo de limpeza, normalização e padronização de dados com técnicas de remoção de símbolos e caracteres especiais, remoção de tags HTML (Hypertext Markup Language), remoção de stop words (preposições, pronomes,

artigos), correção de grafia, stemming (reduzir palavras ao seu radical) e lematização (reduzir a flexão das palavras).

Os itens de normalização utilizados neste trabalho foram:

1. Remoção de tags HTML.
2. Remoção de símbolos e caracteres especiais.
3. Remoção de palavras irrelevantes (Stopwords).
4. Lematização.

Uma vez que os comentários da RHS seguem um padrão de escrita compreensível e coerente, itens como a correção gráfica não se mostra de extrema importância nessa análise, por outro lado, os comentários têm inúmeras tags HTML que foram tratadas durante o processo (Apêndice A).

Por sua vez, seguindo as outras etapas da normalização, como a remoção de caracteres especiais que removem itens como (“”, ?, !, /, ;, :), e a remoção de palavras irrelevantes, que é feita a partir de um dicionário de artigos, preposições, pronomes (a, e, é, ali, uns), entre outros, com isso o texto passa a ser transformado novamente, como é demonstrado no Apêndice B.

Vale mencionar que as stopwords, podem seguir determinados padrões, nessa análise foi usado um padrão comum da língua portuguesa e foram adicionados alguns termos que se encontravam nos comentários e não agregavam valor para a análise final.

A lematização por sua vez, se trata do da remoção de afixos das palavras, fazendo com que essa palavra volte para sua base raiz (SARKAR, 2016), levando em consideração a parte do discurso em que a palavra se encontra, ou seja, caso a palavra se encontre na posição de verbo dentro de um texto, ela será lematizada de uma forma, caso seja um adjetivo será de outra forma e assim por diante.

Uma vez que para a lematização é necessário identificar a que parte do discurso a palavra pertence e após isso lematizar a palavra, é possível fazer o uso de diferentes bibliotecas e ferramentas do python para essa tarefa, neste trabalho foi utilizada a biblioteca Spacy2, que trabalha com uma variedade de idiomas incluindo o Português.

A lematização se faz importante pois, como demonstrado o léxico utilizado tem um número relevante de palavras que são representadas como parte do discurso, como adjetivos e verbos, conseguir lematizar o texto impacta no resultado final de identificação do sentimento de cada comentário.

Para identificar os sentimentos, foi utilizada uma abordagem similar ao utilizado para verificação de sentimentos com o léxico de opinião (HU;LIU, 2004), que se trata de um léxico com palavras positivas (1) e negativas (-1), entranto, o OpLexicon traz palavras neutras (0), ao adaptar o script da biblioteca NLTK3 que faz uso do léxico de opinião foi possível identificar o sentimento dos comentários.

De forma geral, o script acessa cada um dos comentários, separa as palavras, verifica se essa palavra se encontra no dicionário léxico e qual sua polaridade, após esse processo, é feita uma contagem e caso existam mais palavras positivas no comentário, o mesmo é considerado positivo e assim por diante.

4 Resultados

O trabalho com texto gera resultado diversos, que vão além da identificação de sentimentos, dessa forma foi feita uma verificação nos textos e palavras que geraram uma visão descritiva dos dados. Como por exemplo as palavras mais comuns encontradas nesses mais de 36 mil comentários (Apêndice C).

Além das palavras mais comuns é possível visualizar a interação em relação ao número de palavras utilizadas nos comentários e qual seria a média padrão de palavras.

De forma geral, foram encontrados mais de 25 mil comentários que foram classificados como positivos, 4,594 negativos e em torno de 4.692 comentários considerados neutros.

Faz-se importante ressaltar que o resultado final não apresenta indicadores de um modelo de aprendizado de máquina como, accuracy, f1-score, entre outros pelo fato de todo o trabalho ter sido desenvolvido em volta de um léxico de sentimento, o que foi feito em si representa a identificação de sentimentos nos comentários de acordo com palavras que estão presentes no léxico de sentimento, não sua classificação, a partir desse trabalho o treino de um classificador de sentimentos se dá em uma etapa posterior.

5 Considerações Finais

O estudo de identificação de sentimentos nos comentários da rede humaniza SUS, apresenta um resultado que demonstra estar de acordo com aquilo que a rede se propõe, que seria a colaboração entre os usuários em torno de uma temática específica.

Identificar sentimentos positivos em sua grande maioria pode ser visto como um incentivo para que o trabalho de desenvolvimento da rede tenha continuidade, foi possível perceber durante o estudo que o teor dos comentários em sua maioria, são de mensagens de agradecimento, por divulgação de encontros ou iniciativas relacionadas a humanização da saúde no Brasil, gerando um assim um ambiente de aprendizado e gratidão por todos aqueles que estão envolvidos nesse processo.

Os resultados obtidos em relação a esse estudo demonstram uma abordagem inicial em relação a tudo aquilo que pode ser produzido a partir do uso da análise de sentimentos e processamento de linguagem natural. A rede humaniza sus é um ambiente próspero para a possível criação de um léxico de sentimento relacionado a saúde e a humanização da saúde.

Essa mesma abordagem gera caminhos para que, a partir dos resultados desse estudo, seja criado um classificador de sentimentos utilizando métodos supervisionados e aprendizado de máquina, o que daria um contexto mais abrangente para os resultados, bem como para o uso dos dados da RHS.

Por fim, este trabalho agrega um valor experimental para o desenvolvimento de análises de sentimento utilizando o Português brasileiro, com ênfase em um método não supervisionado e com a possibilidade de construção de dados para sua posterior aplicação em algoritmos e classificadores de aprendizagem de máquina.

Referências

DOSCIATTI, M. M.; FERREIRA, E. C. L. P. C. Identificando emoções em textos em português do Brasil usando máquina de vetores de suporte em solução multiclasse.In: **ENIAC-Encontro Nacional de Inteligência Artificial e Computacional.** Fortaleza, Brasil, 2013.

DUARTE, E. S. **Sentiment analysis on twitter for the portuguese language.** Tese (Doutorado) — Faculdade de Ciências e Tecnologia, 2013.

FREITAS, L. d; VIEIRA, R. Exploring resources for sentiment analysis in portuguese language. In: IEEE. In: 2015 **Brazilian Conference on Intelligent Systems**. BRACIS. [S.l.], 2015. p. 152–156.

HU, M; LIU, B. Mining and summarizing customer reviews. **Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, Seattle, Washington, p.168-177, 2004

LIU, B. **Sentiment Analysis and Opinion Mining: Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 180p., 2012.**

SARKAR, D. **Text Analytics with Python**: **A Practical Real-World Approach to Gaining Actionable Insights from Your Data**. Library of Congress Control Number, Apress, 385p., 2016.

SOUZA, M.; VIEIRA, R.; BUSETTI, D.; CHISHMAN, R. e ALVES, I. M. Construction of a Portuguese Opinion Lexicon from multiple resources. In: **8th Brazilian Symposium in Information and Human Language Technology**, 2012.

Apêndice A – Exemplo de remoção de tags HTML.

```
<P>Excelente post, Mariella!<P>
<P>Acho que você e o Bruno poderão e deverão ter um papel
fundamental nesta Rede, orientando-nos em relação a como
redigir um post<P>
<P>Precisaremos, certamente, de um "manual de redação"!</P>
<P>Um abraço,</P>
<P>Ricardo<BR></P>

Excelente post, Mariella! Acho que você e o Bruno poderão e deverão
ter um papel fundamental nesta Rede, orientando-nos em relação a
como redigir um post. Precisaremos, certamente, de um "manual de redação"!
Um abraço, Ricardo
```

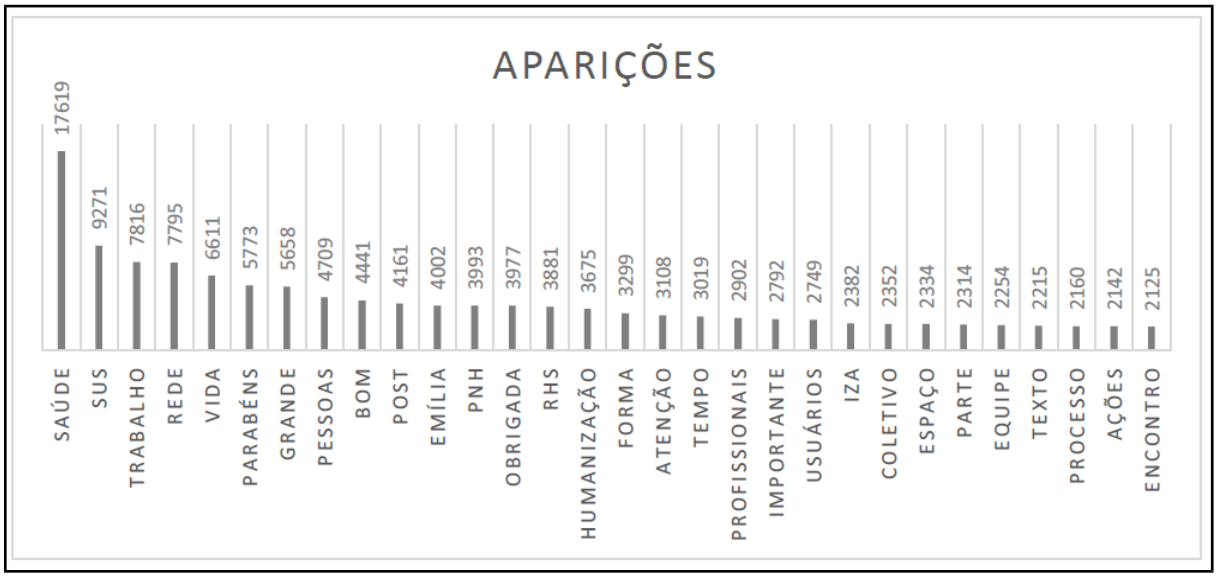
Fonte: Dados da pesquisa, 2018.

Apêndice B – Exemplo de remoção de símbolos e caracteres especiais e stopwords.

```
excelente post mariella acho bruno poderão deverão
papel fundamental rede orientando-nos redigir post
precisaremos certamente manual redação abraço ricardo
```

Fonte: Dados da pesquisa, 2018.

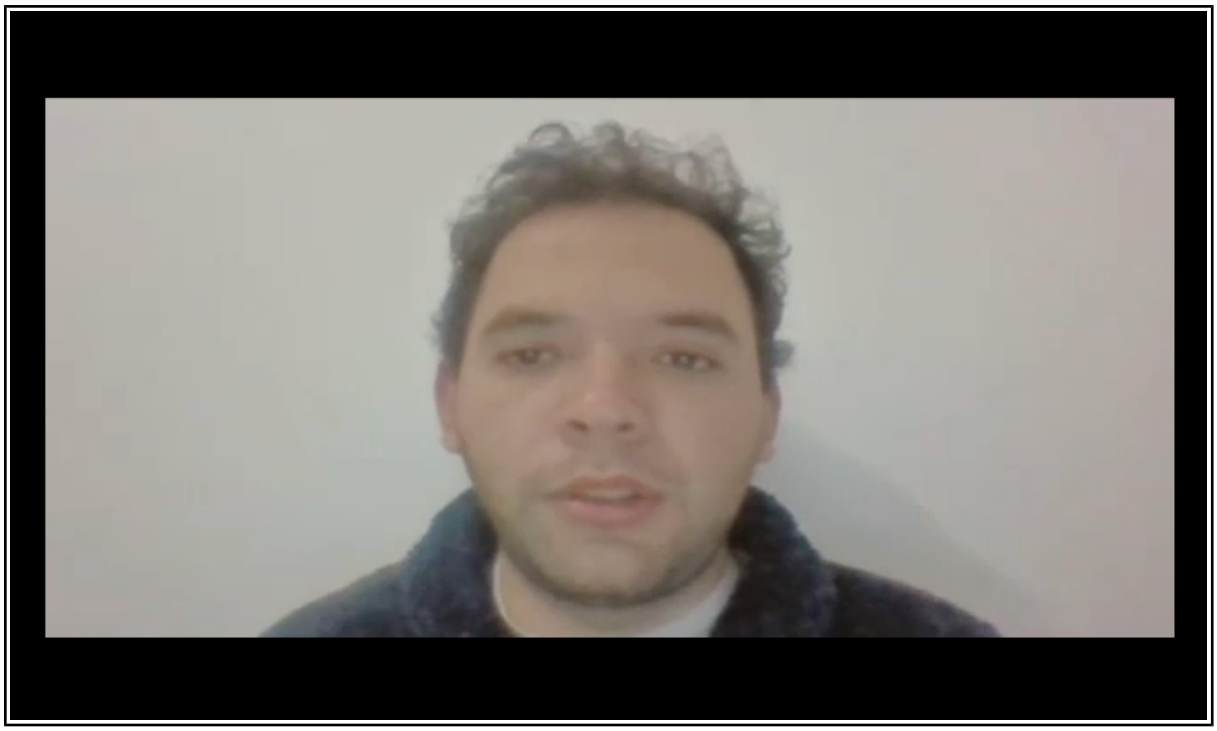
Apêndice C – 30 palavras mais comuns nos comentários.



Fonte: Dados da pesquisa, 2018.

Vídeo da apresentação

Título: Análise de sentimentos: Identificando sentimentos em comentários da Rede Humaniza SUS.



Disponível em: http://dadosabertos.info/enhanced_publications/idt/video.php?id=19

Transcrição da apresentação

O meu é Eduardo esse vídeo é uma apresentação sobre o trabalho que foi enviado para o Workshop de Informação, Dados e Tecnologia, que ocorre na Universidade da Paraíba entre os dias 27/11 e 29/11 de 2018.

O trabalho em questão tem o nome de ANÁLISE DE SENTIMENTOS: Identificando sentimentos em comentários da Rede Humaniza SUS o trabalho foi produzido por mim Eduardo Silva e pelo meu colega Luiz Felipe Rosa.

O trabalho em questão tem como objetivo identificar sentimentos nos comentários da rede humaniza sus é uma rede social que têm atividade desde o ano de 2008 permeia assuntos relacionados ao sistema único de saúde brasileiro, ou seja, o SUS.

Atualmente a rede tem cerca de 30 mil usuários 14 mil publicações e em torno de 35 mil a 40 mil comentários. Para análises de sentimentos da rede dos comentários da rede nós fizemos o uso de mineração de dados ou mineração de texto nesse caso aplicando alguns conceitos de processamento de linguagem natural tudo utilizando a linguagem de programação python os dados foram coletados utilizando um banco de dados da rede após a coleta dos dados nós passamos a um tratamento desses dados usando algumas metodologias de análise de texto e processamento de linguagem natural.

O texto em si vinha com linha com uma sujeira, ou seja, nesse contexto a sujeira são, por exemplo, caracteres assim que não foram bem identificados como a "Ç", o "ÃO" entre outros caracteres.

Para além disso alguns comentários vinham com tags HTML, ou seja, apresentava as tags de parágrafos quebra de linha entre outros utilizando a linguagem de programação python fizemos limpeza desses dados e a sua normalização, ou seja, os dados passaram a ser normalizados sem sujeira todos os comentários com letras minúsculas.

Para identificação de sentimentos ou em alguns casos a sua classificação que não é bem esse caso podemos utilizar o aprendizado de máquina ou um léxico de sentimento uma vez que o intuito do trabalho para identificação e não a classificação nós utilizamos um léxico sentimento, o léxico é um com um dicionário com um conjunto de palavras ou textos que tenha atribuído a essas palavras uma polaridade essa popularidade pode ser positiva negativa ou neutra no caso do estudo nós utilizamos um léxico de sentimento chamado o OpLéxicon que é produzido pela PUC-RS baseado em texto jornalístico e em algumas outras fontes que eles utilizaram para produzir.

O OpLéxicon contém cerca de 32191 itens, ou seja, 32191 palavras dentre elas temos 24485 objetivos e 6889 verbos.

Nesse caso é preponderante o número de adjetivos e verbos uma vez que na língua portuguesa para definir um sentimento normalmente são esses dois tipos de palavras que são utilizados mas o léxico ainda tem hashtags determinantes preposições adjetivos e emoticons para dar segmento à metodologia o que fizemos foi a parte do léxico do sentimento criar uma metodologia de comparação das palavras que existe no comentário e que aparece no léxico, ou seja, se uma sentença tem 20 palavras e 10 das palavras aparecem no léxico de sentimento nós iremos armazenar os valores da polaridade dessas palavras e se encontram tanto no léxico quanto

no comentário em questão a partir disso nós usamos uma metodologia de verificação e cálculo para definir quais os sentimentos aquele comentário que apresenta para tal nós tivemos como base os estudos feitos por autores que criaram um léxico de opinião o único de mineração de opinião da língua inglesa a partir do código que faz o uso desse léxico, nós fizemos uma adaptação com o léxicon a partir disso conseguimos verificar quais eram as popularidades das palavras e assim sendo vai ficar com a popularidade do comentário, ou seja, se o comentário era positivo negativo ou neutro de acordo com o número de palavras positivas negativas ou neutras que apareciam nesse comentário.

Feito isso nós replicamos esse processo para todos os quase 40 mil comentários gerando assim a identificação de sentimentos de cada um deles.

A partir da identificação da popularidade nós podemos perceber que existe uma quantidade maior de comentários positivos e neutros do que negativos.

Talvez isso ocorra por conta do tipo de rede social ao qual estamos lidando é uma rede um pouco mais controlada e focado em um tema específico então as pessoas nos comentários tendem a ser mais assertivas em termos de não dar comentários negativos mas sim comentários positivos de apoio ou então mensagens mais simples como bom dia gostei muito da postagem da publicação e comentários nessa linha

Dessa forma o resultado final que nós tivemos além da popularidade dos sentimentos a verificação das palavras mais comumente utilizadas em todos os comentários sendo que a palavra saúde como esperado parece em sua maioria assim como SUS entre outras relacionadas a esse contexto da saúde.

Foi possível também averiguar algumas questões mais minuciosos que se tratam de uma análise um tanto quanto descritiva como, por exemplo, qual a frequência de comentários por mês por ano e assim por diante. Mas o resultado final a que nos interessava e que foi alcançado era definir ou identificar os possíveis sentimentos de cada um dos comentários pode se dizer que de certa forma nós estamos criando um dataset que posteriormente pode ser utilizado para classificar novos comentários então é importante ressaltar que o trabalho não se trata da classificação utilizando aprendizado de máquina mas sim da identificação utilizando um léxico a partir desse ponto é possível fazer uma comparação com outros léxicos de sentimento da língua portuguesa para verificar quais deles têm uma taxa de acerto maior em relação à identificação de sentimentos, no entanto, uma vez que foi feito uso apenas do OpLéxico.

Nós tivemos como resultado final somente esses dados de identificação do sentimento agradeço a atenção daqueles que ler um artigo ou que estão apenas vendo o vídeo de apresentação muito obrigado e até mais.