UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS DA SAÚDE
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA EM SAÚDE
LINHA DE PESQUISA: TECNOLOGIA DE INFORMAÇÃO E COMUNICAÇÃO EM SAÚDE/ESAÚDE

João Augusto da Silva Bueno

MODELO CLASSIFICADOR AUTOMÁTICO DE POLARIDADE DE OPINIÕES NO FACEBOOK®

Florianópolis
2019

João Augusto da Silva Bueno

MODELO CLASSIFICADOR AUTOMÁTICO DE POLARIDADE DE OPINIÕES NO FACEBOOK®

Dissertação Mestrado Profissional em Informática em Saúde do Centro de Ciências da Saúde da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do Título de Mestre em Informática em Saúde.

Florianópolis
2019

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Bueno, João Augusto da Silva
Modelo classificador automático de polaridade de opiniões no Facebook© / João Augusto da Silva Bueno ; orientador, Grace Teresinha Marcon Dal Sasso, 2019.
86 p.

Dissertação (mestrado profissional) - Universidade Federal de Santa Catarina, Centro de Ciências da Saúde, Programa de Pós-Graduação em Informática em Saúde, Florianópolis, 2019.

Inclui referências.

1. Informática em Saúde. 2. Análise de sentimentos. 3. Mineração de texto. 4. Aprendizagem de máquina. 5. Câncer de próstata. I. Teresinha Marcon Dal Sasso, Grace. II. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Informática em Saúde. III. Titulo.

João Augusto da Silva Bueno

MODELO CLASSIFICADOR AUTOMÁTICO DE POLARIDADE DE OPINIÕES NO FACEBOOK®

O presente trabalho em nível de mestrado profissional foi avaliado e aprovado por banca examinadora composta dos seguintes membros:

Grace Teresinha Marcon Dal Sasso:54406854991
Assinado de forma digital por Grace Teresinha Marcon Dal Sasso:54406854991
Dados: 2019.10.14 11:48:14 -03'00'

Prof.ª Dr.ª Grace Teresinha Marcon Dal Sasso
Orientadora
Universidade Federal de Santa Catarina

Profª Dra. Gabriela Marcellino de Melo Lanzoni
Universidade Federal de Santa Catarina

Profº Dr. Paulino Sousa
Escola Superior de Enfermagem do Porto

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Informática em Saúde.

Grace Teresinha Marcon Dal Sasso:54406854991
Assinado de forma digital por Grace Teresinha Marcon Dal Sasso:54406854991
Dados: 2019.10.14 11:47:52 -03'00'

Prof.ª Dr.ª Grace Teresinha Marcon Dal Sasso
Coordenadora do Programa

Prof.ª Dr.ª Grace Teresinha Marcon Dal Sasso
Orientadora

Florianópolis, 30 de setembro de 2019.

A Deus e à minha família.

AGRADECIMENTOS

Primeiramente agradeço a Deus, pela vida, por nunca me deixar desistir dos meus sonhos, mesmo aparecendo muitas indecisões e dificuldades pelo caminho.

Aos meus pais, Jucelino e Graciéla, pelos ensinamentos, conselhos, pelos bons exemplos e principalmente pelo apoio em todas as situações e o amor envolvido.

A meu irmão, André, por me apoiar nas horas de dificuldades.

À minha esposa, Liana, por dividir, acreditar, apoiar e participar ativamente dos meus sonhos, me dando suporte suficiente para aguentar a pressão do cotidiano.

A meus filhos, Ariele e José Augusto, que são a razão de todos os meus esforços, que talvez sem saber sejam o combustível por tudo que faço em minha vida.

A todos os meus familiares, por entender, respeitar e colaborar nessa etapa da minha vida.

Aos amigos, que entenderam os "nãos", os "não vou poder ir", os "estou apurado".

Aos companheiros do IFSC Caçador, a todos os servidores, direção, os técnicos, até o pessoal terceirizado, que me receberam muito bem naquela cidade e depois me incentivaram e me apoiaram no inicio dessa jornada. Em especial aos professores, principalmente ao pessoal da área de informática, professores André, Egon, Samuel e Tiago.

Aos colegas do IFSC Lages, a todos os servidores, direção, os técnicos até o pessoal terceirizado, aos amigos professores da área de informática, Juliano, André, Vilson Heck, Perin, Wilson Castelo Branco, Robson, José Sé, Roberval e Thaiana, que colaboraram com dicas e apoio para que o meu afastamento parcial pudesse acontecer.

A meu amigo e professor Davi, pelos incentivos, dicas e conselhos sobre o mestrado, amigo que nunca mediu esforços para me ajudar trocando o horário das suas aulas comigo para que eu pudesse viajar para as aulas.

A meu amigo Valter, por ter me incentivado a estudar, buscar uma profissão, iniciando a minha jornada na informática, no ano de 2001.

A meus amigos da Escolas Sid.

A meu sogro e amigo Evandro, companheiro de faculdade.

A meu amigo Wandrey da UTFPR, você foi muito importante!

A meu amigo e professor Alex, que é meu parceiro de estudos, de trabalhos e amigo para todas as horas.

À minha amiga Márcia Sagaz, que me apoiou muito nessa etapa, seja com conselhos, com broncas ou com a sua experiência em trabalhos. Te agradeço muito! Obrigado por tudo!

A todos meus amigos do Programa de Mestrado em Informática em Saúde, com certeza sem vocês eu não chegaria nesse momento. Obrigado a todos pelo grande apoio, posso dizer que fiz grandes amigos nesse tempo. Muitas histórias para contar, estudos, rizadas, aflições, caronas, que saudade. Obrigado pessoal.

A todos os professores do Programa de Mestrado em Informática em Saúde. Obrigado pelos ensinamentos.

À minha orientadora Professora Grace, pelas dicas, ensinamentos, orientações e principalmente pelos conselhos na etapa final do percurso. Obrigado mesmo.

À UFSC pela grande oportunidade de me tornar um profissional mais qualificado e aumentar os meus conhecimentos.

RESUMO

As campanhas de prevenção a doenças, principalmente as realizadas na internet, são de suma importância nos dias atuais, pois por meio delas é possível entender, em tempo real, ao menos parcialmente, se a campanha está sendo bem recebida pela população. Para esse monitoramento é necessário acompanhamento automático, que trabalhe como ferramenta auxiliadora dos gestores nos processos de tomada de decisão. A proposta deste estudo é desenvolver e analisar um modelo classificador automático de polaridade das opiniões, aplicado às postagens sobre o câncer de próstata da página denominada Novembro Azul no Facebook®, no período de novembro de 2018. O presente estudo é uma produção tecnológica inovadora de natureza quantitativa. Foi desenvolvido em cinco etapas, primeiramente, foram extraídos os dados da rede social para construção de um *dataset* em uma planilha eletrônica. Logo depois, na segunda etapa, foi realizado o pré-processamento dos dados, retirando os sufixos das palavras, removendo as palavras repetidas ou sem relevância na formação das frases. Em seguida, uma amostragem dos dados recebeu classificação: positiva ou negativa. Essa ordenação chamada de treinamento foi realizada por três especialistas em tecnologia. A quarta etapa foi voltada para a criação e utilização do algoritmo de aprendizagem de máquina, denominado *naive bayes*, que analisou e processou o treinamento que os classificadores especialistas realizaram na etapa anterior, fazendo a classificação de forma automática do restante do material disponível na base de dados. Na sequência, foi efetivada a análise do modelo, medida a sua acurácia e aplicados métodos para auxiliar na visualização dos resultados, podendo assim ser realizadas alterações para ajustes no código fonte. O modelo obteve no teste da acurácia o resultado de 84,8% na classificação automática das opiniões em relação à classificação dos especialistas, concluindo assim que o modelo implementado poderá auxiliar os gestores das campanhas de saúde nos processos de tomada de decisão.

Palavras-chave: Mineração de texto. Análise de sentimentos. Aprendizagem de máquina. Câncer de próstata. Redes sociais.

ABSTRACT

Disease prevention campaigns, especially those carried out on the Internet, are of paramount importance today, because through them it is possible to understand, in real time, at least partially, if the campaign is being well received by the population. This requires automatic monitoring, which works as an auxiliary tool for managers in decision-making processes. The purpose of this study is to develop and analyze an automatic opinion polarity classifier model applied to the postings on prostate cancer on the Facebook® page named November Blue, in November 2018. This study is an innovative technological production of quantitative nature. It was developed in five steps. First, we extracted data from the social network to build a dataset in an electronic spreadsheet. Soon after, in the second stage, the data was preprocessed, removing the suffixes of the words, the words repeated or not relevant in the formation of the sentences. Then a sampling of the data was graded: positive or negative. This ordination named training was performed by three technology experts. The fourth step focused on the creation and use of the machine learning algorithm, called *naive bayes*, which analyzed and processed the training that the expert classifiers performed in the previous step, automatically classifying the remaining material available in the database. Subsequently, the model analysis was performed, its accuracy was measured and methods were applied to assist in the visualization of the results, thus making changes to adjustments to the source code. The model obtained in the accuracy test the result of 84.8% in the automatic classification of opinions in relation to the experts' classification, thus concluding that the implemented model may help health campaign managers in decision-making processes.

Keywords: Text Mining. Sentiment analysis. Machine learning. Prostate cancer. Social networks.

LISTA DE SIGLAS

AM	Aprendizagem de Máquina
API	Programação de Aplicativo
BD	Banco de Dados
CSV	Comma-Separated Values
FD	Funções Descritivas
FP	Funções Preditivas
IA	Inteligência Artificial
IC	Inteligência Computacional
INCA	Instituto Nacional de Câncer
RI	Recuperação da Informação
SBU	Sociedade Brasileira de Urologia
SL	Aprendizagem Supervisionada
MT	Mineração de Texto
PLN	Processamento de Linguagem Natural
UICC	International Union Against Cancer

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

O crescimento das redes sociais proporcionou oportunidades de pesquisa para rastrear comportamentos públicos, informações e opiniões sobre muitos problemas, inclusive os comuns na área da saúde. Estima-se que o número de usuários de redes sociais aumentará de 2,34 bilhões em 2016 para 2,95 bilhões em 2020 (STATISTA, 2017).

Com a expansão dessas redes, a quantidade de dados na *web* também aumentou, assim como a forma com que as pessoas procuram conhecimento. A relevância do que é postado se modificou, pois além dos conteúdos disponibilizados, os usuários buscam nas redes sociais, comentários e posições sobre diversos assuntos e temas (DE ARAÚJO, 2012).

Contudo, trabalhar com um grande volume de postagens ou publicações é uma tarefa complexa, principalmente por serem dados do tipo não estruturados, isto é, com escrita livre. Esses dados, em muitas ocasiões, incluem informações importantes, como: tendências, anomalias e padrões de comportamento que podem ser utilizados para apoiar tomadas de decisões de modo geral (BERRY; KOGAN, 2010). Ainda, grandes empresas como Google, Amazon e Bloomberg investigam sobre a análise de big data usando classificação de tópicos e mineração de mídia social sobre o que as pessoas estão dizendo e selecionam dados relevantes por cluster ou outros métodos de aprendizado de máquina. Depois que os clusters de interesse são identificados, então, a aplicação das técnicas de processamento de linguagem natural é uma das vias mais promissoras para o processamento de dados de mídia social (FARZINDAR; INKPEN, 2015).

Diante disso, faz-se necessária a utilização de técnicas de extração, manipulação e tratamento dos dados, analisando-os de forma consistente. Essa técnica, conhecida como mineração de dados, tem grande importância para identificar e/ou acompanhar comportamentos, sendo utilizada em pesquisas de diversas áreas, seja estudos acadêmicos ou levantamentos para gestão de *marketing* de empresas. Outro recurso empregado para examinar mensagens das redes sociais é a analise de sentimentos, ela utiliza muitos métodos, técnicas e ferramentas para detectar e extrair informações subjetivas, como opiniões da linguagem utilizada pelos usuários (MÄNTYLÄ *et al*., 2019). Existem na literatura muitos trabalhos em que a mineração de dados é empregada para análise de sentimentos, principalmente para analisar opinião sobre produtos e serviços (GIATSOGLOU *et al*., 2017). Entre esses estudos, estão, por exemplo, a análise de comentários de filmes (OHANA; TIERNEY, 2009), a de discursos políticos (THOMAS; PANG; LEE, 2006), a análise de produtos, buscando a fraqueza destes, a partir das opiniões dos clientes (ZHANG; XU; WAN,

2012); um exemplo concreto é o caso da empresa IBM® que monitorou no Twitter® o tema futebol, buscando saber o que os torcedores estavam falando dos seus times (EXAME, 2014), ainda, o intercâmbio de informações financeiras com os usuários do Facebook® e Twitter® ou pesquisa sobre intenções de voto baseados nas mensagens de mídia social (Farzindar; Inkpen, (2015). Outro exemplo é o de universidades que podem se interessar em verificar a aceitação ou não de um novo curso, monitorando as opiniões em relação a esse tema em uma rede social (GONÇALVES *et al*., 2013).

Dessa forma, a utilização das redes sociais como fonte de pesquisa para a análise de opiniões abre possibilidades com a busca e a mineração dessas opiniões (LIU, 2010).

Sob outro ângulo, a extração inteligente de dados das redes sociais atraiu interesse em muitos campos, e, a comunidade de Informática em Saúde também lançou mão dessa extração para melhorar simultaneamente os resultados de saúde usando, neste caso as opiniões geradas pelo consumidor/usuário dos serviços de saúde (GHORPADE, 2017).

Recentemente, houve um rápido aumento no interesse em relação à análise de redes sociais em comunidades de mineração de dados. A motivação básica é a demanda para explorar o conhecimento a partir de grandes volumes de dados coletados, referentes a opiniões de usuários em ambientes *on-line*. As técnicas baseadas em mineração de dados estão se mostrando úteis para a análise de dados de redes sociais, especialmente para grandes conjuntos de dados que não podem ser manipulados por métodos tradicionais (INJADAT; SALO; NASSIF, 2017). Com o crescimento da mineração de dados (por exemplo, revisões, discussões em fóruns, *blogs* e rede social) na *web*, indivíduos e organizações estão usando cada vez mais a opinião pública nesses meios para a tomada de decisões. Os clientes em potencial também querem saber a opinião dos usuários antes de usar um serviço ou tratamento, aderir a um programa de saúde ou comprar um produto (GHORPADE, 2017).

A mineração desses dados, também conhecida como descoberta de conhecimento em um banco de dados é um desenvolvimento que serve para acessar e extrair informações desse banco (CHAN *et al*., 2011). A mineração de dados aplica técnicas de aprendizado de máquina e análise estatística para a descoberta automática de padrões em um banco de dados. Já a mineração de texto é um domínio especializado que aplica técnicas de mineração de dados ao texto, enquanto a análise de sentimentos visa a identificar sentimentos manifestados por meio das opiniões dos indivíduos e comunidades (ISAH; TRUNDLE; NEAGU, 2014).

Entende-se que o conceito de opinião, por vezes pode se confundir com o de sentimento e emoção e até mesmo de mensagem, por isso, elaborou-se a seguinte figura com o objetivo de apresentar de maneira mais clara cada um desses elementos. Sendo a mensagem

o veículo que carrega ou contém a opinião que por sua vez é constituída de sentimento(s) que resulta(m) em emoção.

As técnicas de mineração de texto podem ser aplicadas para investigar a opinião do consumidor em relação a marcas globais, e algumas redes sociais já foram utilizadas como locais confiáveis para a análise de opiniões (MOSTAFA, 2013). Nesse sentido, a mineração de opinião em fóruns médicos é semelhante à mineração de opinião em *blogs*, mas o interesse está focado em um problema de saúde específico. Uma etapa preliminar a esse processo pode ser a filtragem automática das postagens para manter apenas as relevantes ao tópico (e incluir opiniões). Para a tarefa de classificação de opinião, as técnicas são baseadas em aprendizado de máquina e / ou contagem de polaridade de termos (FARZINDAR; INKPEN, 2015).

Dessa forma, a análise de dados de redes sociais pode colaborar com as organizações de saúde, departamentos de saúde e sistemas de saúde, fornecendo indicadores que podem ser eficazes no planejamento e análise das cidades, estados e países (MEJOVA *et al*., 2015).

Os profissionais de saúde podem usar a opinião do próprio usuário para melhorar seus serviços. Os médicos podem coletar *feedback* de outros médicos e de pacientes para melhorar suas recomendações e resultados de tratamento. O consumidor pode usar o conhecimento de outros consumidores para tomar decisões sobre saúde.

Nesse sentido, Ghorpade (2017) ressalta que as redes sociais identificam grupos e influências dos usuários resultando na geração de grandes quantidades de contribuições textuais, apoiando as decisões governamentais, coletando opiniões relevantes e conhecimentos valiosos da sociedade, no entanto, são explorados de forma limitada. É fundamental analisar essas colaborações de forma minuciosa (MARAGOUDAKIS; LOUKIS; CHARALABIDIS, 2011).

Ainda, os debates relacionados à saúde nas redes sociais podem incluir informações geradas por usuários que podem fornecer informações sobre problemas de saúde pública, como sintomas, abuso de medicações, reações adversas, efeitos em longo prazo e a utilização de múltiplas drogas (SARKER; GONZALEZ 2016).

Muitas dessas contribuições são realizadas a partir das redes sociais, e de modo interessante em números muito mais altos do que nos debates de política pública. Em geral, contêm opiniões sobre a política, decisões do governo em discussão e conhecimento sobre as necessidades sociais da população, como saúde, educação, entre outros (MARAGOUDAKIS; LOUKIS; CHARALABIDIS, 2011).

Entre as diversas ações que envolvem a área da saúde, estão as campanhas de prevenção e entre elas a do câncer de próstata, sendo esse o tumor é o mais frequente no homem, com 68 mil novos casos previstos para 2019 (INCA, 2018).

Atualmente, a *fanpage* denominada Novembro Azul[1] participa da campanha Novembro Azul, que é a maior campanha de combate ao câncer de próstata no Brasil. Essa *fanpage* recebe muitos comentários e opiniões sobre as suas publicações e tem sido referência na missão de orientar a população masculina a cuidar melhor da própria saúde (LADO A LADO PELA VIDA, 2019).

Em relação aos dados, com intuito de facilitar a sua extração em redes sociais, várias organizações e empresas, incluindo plataformas de mineração de textos, recentemente, criaram e disponibilizaram a chamada interface de programação de aplicativos (API) (do inglês *Application Programming Interface*) para compartilhar os dados públicos dos seus grupos, de suas páginas e de seus usuários. No estágio de coleta e limpeza de texto, uma chamada da API para a autenticação e a extração de dados é solicitada no Facebook®. A API do Facebook® é usada para buscar páginas, atualizações de *status* e comentários que sugiram com as experiências e visualizações dos usuários (ISAH; TRUNDLE; NEAGU, 2014).

Assim, neste trabalho utilizou-se a API do Facebook® na *fanpage* denominada Novembro Azul, que atualmente conta com mais de 128.000 seguidores, e é administrada pelo Instituto Lado a Lado pela Vida, que trabalha com a multiplicação de informações sobre saúde cardiovascular e câncer para transformar a vida das pessoas por meio de campanhas de prevenção, diagnóstico e tratamento. Sua missão é empoderar o brasileiro, e em 2008, criou o movimento "Um Toque, um Drible" que deu origem a *fanpage* Novembro Azul, em 2012 (LADO A LADO PELA VIDA, 2019).

Assim, neste estudo, mineração de opinião é definido como a tarefa de detectar, extrair e classificar opiniões sobre alguma coisa. É um tipo de processamento da linguagem natural (PNL) para rastrear o humor do público em relação a uma determinada lei, política ou *marketing* etc. Envolve uma maneira de desenvolvimento para a coleta e análise de comentários e opiniões sobre legislação, leis, políticas, etc., publicadas nas mídias sociais.

O processo de extração de informações é muito importante porque é uma técnica útil, mas também uma tarefa desafiadora. Isso significa que, para extrair opiniões de um objeto na internet, é necessário automatizar sistemas de mineração de opinião para fazê-lo. Dentre as

[1] Disponível em: https://www.Facebook.com/NovembroAzulBrasil/. Acesso em: 7 jul. 2019.

técnicas existentes para análise de sentimentos podemos salientar o aprendizado de máquina (supervisionada e não supervisionadas) e abordagens lexicais. (SABERI; SAAD, 2017)

Fundamentado no alto número de informações contidas nas redes sociais, apoiado pela evolução das técnicas de mineração de dados e análise de sentimentos, embasado nos dados epidemiológicos sobre a quantidade de homens afetados com o câncer de próstata no Brasil; respaldado pela importância das campanhas de saúde; pautado na relevância da busca, mineração e análise dos dados de textos das redes sociais, considera-se importante propor um estudo que busque e utilize um método que procure dados, faça a mineração, classifique e apresente esses dados de forma que possam auxiliar a gestão em saúde, que sejam capazes de monitorar a campanha de prevenção ao câncer de próstata realizada na *fanpage* Novembro Azul, no Facebook®, e, assim poder ser um *feedback* de como a sociedade está aderindo às ações de prevenção ao câncer de próstata a partir da campanha veiculada. Diante desse contexto, este estudo busca responder a seguinte questão: que ferramenta pode ser utilizada para classificar as opiniões manifestadas por meio de texto livre pelos seguidores da página do Facebook® denominada Novembro Azul relacionada ao câncer de próstata?

1.1 OBJETIVOS

Diante do contexto apresentado, este estudo tem como objetivo geral desenvolver e analisar um modelo classificador automático de polaridade das opiniões manifestadas por meio de postagens dos seguidores na campanha realizada em novembro de 2018, sobre o câncer de próstata, pela página denominada Novembro Azul no Facebook®. Para alcançá-lo foram estabelecidos os seguintes objetivos específicos:

- Verificar na literatura a(s) ferramenta(s) e/ou modelo(s) de extração, análise e classificação automática de textos não estruturados em redes sociais.
- Criar e aplicar o modelo de extração, pré-processamento, treinamento, classificação e visualização automatizada nos textos não estruturados da campanha de prevenção ao câncer de próstata.
- Medir o grau de acurácia do modelo classificador automático de manifestações das emoções.

O presente estudo está organizado em sete capítulos, para além desta introdução. O capítulo 2 apresenta a revisão da literatura e conceito, como, redes sociais e mídias sociais, mineração de dados e textos, processamento de linguagem natural e a definição de análise de sentimentos No capítulo 3, é tratada a metodologia do estudo, sua natureza, seu tipo e seu protocolo, também são descritas todas as etapas relativas à realização do estudo. O capítulo 4 aborda os aspectos éticos do estudo. No capítulo 5, são abordadas as etapas de desenvolvimento do método proposto. O capítulo 6 apresenta os resultados e suas análises. Por fim, no capítulo 7, apresentam-se as considerações finais do trabalho.

2 REVISÃO DA LITERATURA

Neste capítulo, apresenta-se a descrição de estudos que utilizam coleta de dados em redes sociais para compreender aspectos da realidade de usuários ou grupos de usuários acerca de diversas áreas de interesse, sendo alguns em casos específicos de pesquisas em informática em saúde.

A escolha dos trabalhos listados neste capítulo foi realizada de forma livre, por meio do acesso de variados repositórios, dentre os quais estão: Scholar, Web of Science e IEEExplore. No processo de busca foram utilizadas estas diferentes combinações de palavras-chave e ou descritores: “Sentiment Analysis”, “Text Mining”, “Opinion Mining in Marketing”, “Health Informatics”, “Social Network” e “Análise de Sentimentos”. Outros trabalhos foram obtidos por meio da análise de citações e referências contidas em artigos, teses, livros e relatórios técnicos.

Atualmente, pode-se encontrar na literatura um número expressivo de trabalhos relacionados à mineração de textos em redes sociais. Para este estudo, apresenta-se sucintamente, textos produzidos na última década, que utilizaram a plataforma Twitter®, Instagram®, Wikis e Facebook®. O fio condutor entre todos é a mineração de textos.

No ano de 2010, Chew e Eysenbach criaram no Canadá um modelo que analisava o conteúdo de mensagens postadas no Twitter® durante o surto da gripe H1N1 em 2009. O estudo mostrou principalmente que as mensagens ajudaram a espalhar a informação sobre os sintomas e, principalmente, a descobrir quais foram as localidades com maior foco da doença, auxiliando assim o poder público na busca de soluções em tempo real.

Em 2011, Cheong e Lee ofereceram, na Austrália, um modelo de análise de sentimentos que fornecia visualizações gráficas sobre potenciais cenários de terrorismo com base nos dados do sentimento público coletados no Twitter®.

No mesmo ano, Gomide *et al*. (2011criaram, no Brasil, uma metodologia de vigilância ativa no Twitter® baseada em quatro dimensões: volume, localização, tempo e percepção do público. Exploraram a dimensão da percepção pública, realizaram uma análise de sentimentos. Essa análise permitiu filtrar conteúdos relevantes e não relevantes para a vigilância da dengue em tempo real.

Em um estudo de caso, Yates e Paquette (2011) exploraram alguns Wikis para o compartilhamento da informação e seus impactos no processo de tomada de decisões eficazes durante o terremoto de 2010 no Haiti. Concluíram que quando empregados adequadamente,

os benefícios do suporte de mídia social são ciclos de decisão mais rápidos e recursos de conhecimento mais completos.

Wang *et al*., em 2012, propuseram um estudo que realiza um processamento de dados em tempo real com mensagens do Twitter®, modelo de sentimento estatístico que avaliaram as mudanças do sentimento público em resposta aos acontecimentos políticos nas eleições de 2012 e as notícias à medida que acontecem. Com base nos dados, o classificador executou com precisão de 59% quatro tipos de sentimentos: negativo, positivo, neutro ou inseguro.

Araujo, por sua vez, em 2014, criou um método de classificação de sentimento, aplicado em mensagens do Twitter®, sobre os temas câncer e diabetes, buscando saber a caracterização da popularidade e a repercussão do tema. Encontrando bons resultados com o seu classificador e indicando a necessidade de melhoras no método.

Em outro estudo, os dados do Twitter® foram testados e avaliados, buscando saber se poderiam colaborar com o gerenciamento de desastres. A pesquisa demonstrou a viabilidade do uso desses dados como parte do processo de gerenciamento de desastres para planejamento e aviso prévio. Para esse estudo foram avaliados os alertas de tsunami em Padang Indonésia e as reações entre os usuários examinados por Carley *et al*. (2016).

Já Crannell *et al*., também em 2016, investigaram comportamentos de uso do Twitter® de pacientes com câncer., medindo a sua felicidade média. O estudo buscou investigar o quanto os pacientes de câncer publicam sobre os estágios da doença. A partir da ideia de hedonométrica (*hedonometric*) aplicada em sistemas como o "Hedonometer (que indica distâncias percorridas)". O sistema *Hedonometer* foi desenvolvido pelo departamento Vermont Complex System Center, da Vermont University e trabalha com análises quantitativas sobre a felicidade, isto é, usando palavras que foram *ranqueadas* a partir de dados coletados do: Google Books®, artigos do New York Times, Music Lyrics e mensagens do Twitter®, apoiado na classificação *crowdsourced*[2], pela Mazon's Mechanical Turk, que realizaram a contagem do valor de cada termo em uma escala de (1) *sad* (triste) e (9) *happy* (feliz). O que foi levantado como hipótese desse estudo foi que quanto mais avançado o estágio da doença, maior o número de publicações dos pacientes, e que a felicidade varia entre cada fase da doença diagnosticada. Os autores descobriram que pacientes com câncer descrevem e explicam seus sentimentos sobre suas doenças de forma aberta e franca no Twitter®. Este estudo se realizou com base em 36 milhões de *tweets* coletados.

[2] Em tradução livre: colaboração coletiva

Em outro estudo, Park *et al*. (2016) investigaram a propagação das questões discutidas durante os debates na televisão na eleição presidencial sul-coreana de 2012 para conversas no Twitter® e os padrões de comunicação dos seus usuários em discussões políticas. De acordo com os resultados, as redes de questões tendem a evoluir de acordo com as tendências da audiência do debate na TV, sugerindo que o consumo da mídia de massa para fins políticos está correlacionado com a participação política nas mídias sociais. Esses dados confirmam os efeitos de propagação da mídia tradicional nas mídias sociais.

Em 2016, Rodrigues elaborou um trabalho no qual desenvolveu uma ferramenta de análise de sentimento chamada SentiHealth-Cancer (SHC), cujo objetivo foi auxiliar a detecção do estado emocional de pessoas membros de comunidades virtuais do Brasil que visam ao apoio a pacientes com câncer. Para isto foi realizado estudo comparativo entre a ferramenta proposta e outras quatro ferramentas de propósito geral de análise de sentimento que coletaram 789 mensagens de oito comunidades do Facebook®, além disso, foram levadas em consideração 2.574 avaliações de voluntários sobre os sentimentos expressos nas mensagens coletadas. Para testar os desempenhos das ferramentas em cada uma das oito comunidades, o estudo lançou mão de avaliações de psicólogos e de não psicólogos e de textos traduzidos do português para o inglês, sempre que a postagem era legível suficiente, livre de gírias, de abreviaturas etc.

Como resultado, Rodrigues (2016) obteve que o desempenho do método proposto em seu estudo é superior aos que serviram de comparação, tanto quando analisam textos em português quanto em inglês. Isso se comprova por meio de seu valor de acurácia (56.64%) em relação às demais ferramentas que apresentam acurácia (50.67%), ou seja, um aumento de 5.97% em relação à maior apresentada pelas outras ferramentas.

No ano de 2017, Reece e Danforth usaram ferramentas de aprendizado de máquina para identificar marcadores de depressão. Os recursos estatísticos foram computacionalmente extraídos de 43.950 fotos de participantes do Instagram, usando análise de cores, componentes de metadados e detecção algorítmica de faces. Os modelos resultantes superaram a taxa média de sucesso de diagnóstico não assistido dos clínicos gerais para depressão. Esses resultados se mantiveram mesmo quando a análise foi restrita a postagens realizadas antes que os indivíduos deprimidos fossem diagnosticados pela primeira vez, pois classificações humanas de atributos fotográficos (rostos felizes, tristes, etc.) eram preditores fracos para indícios de depressão e não eram correlacionados com recursos gerados por computador. Esses resultados sugerem novos caminhos para o rastreamento e a detecção precoce de doenças mentais.

No mesmo ano (2017), Dai e Hao realizaram um estudo sobre os danos e benefícios dos cigarros eletrônicos, cujo uso aumentou rapidamente nos últimos anos. Separaram os *tweets* não comerciais, dos comerciais, procuraram avaliar as atitudes do público em geral em relação aos e-cigarros. Coletaram tweets contendo as palavras "e-cig", "e-cigarette", "e-liquid", "vape", "vaping", "vapor" e "vaporizer" publicados de 23 de julho a 14 de outubro de 2015. Como resultado das buscas, foram coletados 757.167 *tweets*. Um modelo multilíngue *naive bayes* foi construído para classificar *tweets* em cinco polaridades (contra, a favor, neutro, comercial, irrelevante). Analisaram ainda a prevalência de tweets de e-cigarros, as variações geográficas desses *tweets* e o impacto de fatores socioeconômicos nas atitudes do público em relação aos e-cigarros. Concluíram que os *tweets* não comerciais aumentaram a conscientização pública sobre os riscos potenciais para a saúde, e podem ajudar a evitar que adolescentes ou adultos jovens usem cigarros eletrônicos. As polaridades de opinião sobre os e-cigarros das redes sociais poderiam influenciar bastante o público em geral, especialmente os jovens. Campanhas educacionais adicionais devem incluir a medição de sua eficácia.

Também em 2017, os autores Hu, Chen e Chou apresentaram trabalho em que foi realizada a mineração da opinião de avaliações de hotéis *on-line*, isto é, o estudo propôs uma técnica para identificar as frases mais informativas de avaliações no Facebook®.

Como se afirmou nesta introdução são diversas as áreas que utilizam dos recursos da informática para propor soluções nesta contemporaneidade. Aqui, dos 13 trabalhos citados, sete referem-se à saúde, sendo os demais distribuídos em política, turismo e catástrofes.

Esta revisão configura tão somente uma amostra da relevância de estudos nesta área, considerando que a Era Digital já é uma realidade. A seguir apresentam-se os conceitos que orientaram o pesquisador a desenvolver o trabalho.

2.1 O CÂNCER E A CAMPANHA NOVEMBRO AZUL

Entre as doenças que mais preocupam as autoridades em saúde está o câncer, que se tornou uma das principais causas de morte em muitos países, sendo que a maneira mais eficaz de reduzir o número de óbitos ocasionados pela doença é descobri-lo mais cedo (KHARYA, 2012). Essa enfermidade é responsável por mais de 12% de todas as causas de óbito no mundo: mais de 7 milhões de pessoas morrem anualmente da doença. Como a expectativa de vida no planeta tem melhorado gradativamente, a incidência de câncer, estimada em 2002 em 11 milhões de casos novos, alcançará mais de 15 milhões em 2020. Essa previsão foi realizada em 2005 pela International Union Against Cancer (UICC) (INCA, 2018).

O câncer de próstata é o mais recorrente homens em todas as regiões do Brasil (70,54 casos novos a cada cem mil indivíduos), com exceção do melanoma, câncer de pele que atinge ambos os sexos. Modesto *et al*. (2018) informam que, o câncer de próstata, "No ano de 2013, foi a segunda causa de mortalidade por neoplasia no sexo masculino, com 14,06 óbitos por cem mil homens, atrás apenas do câncer de traqueia, brônquios e/ou pulmões, com 16,12 óbitos a cada cem mil homens". A maior incidência está entre homens com idade a partir de 65 anos, sendo que esse fator apresenta cerca de 62% dos casos diagnosticados mundialmente. Outros fatores são: história familiar e pele negra. Esse segundo motivo pode, contudo, ser associado a estilo de vida. Outro elemento que pode resultar em risco ou proteção é a dieta do indivíduo.

Sendo a idade o principal fator de risco, é recomendado que o início da prevenção ocorra a partir dos 45 anos, desde que o indivíduo não apresente casos de câncer de próstata na família. Para aqueles que apresentam histórico familiar e para afro-americanos, o acompanhamento deve iniciar aos 40 anos de idade (MEDEIROS; MENEZES; NAPOLEÃO, 2011).

A próstata é uma glândula presente somente nos homens, localizada na frente do reto, abaixo da bexiga, envolvendo a parte superior da uretra (canal por onde passa a urina). A próstata não é responsável pela ereção nem pelo orgasmo. Sua função é produzir um líquido que compõe parte do sêmen, que nutre e protege os espermatozoides. Nos homens jovens, a próstata possui o tamanho de uma ameixa, mas seu tamanho aumenta com o avançar da idade (MINISTÉRIO DA SAÚDE, 2019).

O mau funcionamento da glândula prostática pode explicar a progressão do câncer de próstata. Sendo assim, o não controle da glândula prostática sobre a regulação do hormônio androgênico intraprostático (testosterona) permite com que o câncer cresça, sendo regido a partir dos fatores hormonais de modo autônomo (CAIRE, 2012).

O câncer de próstata é uma patologia que pode ser descoberta precocemente por métodos diagnósticos de triagem. Os exames utilizados para o diagnóstico de câncer de próstata compreendem o antígeno prostático específico (PSA); digital (toque retal); tomografia computadorizada ou ressonância magnética e ultrassonografia transretal (MENEZES, 2017). Diante desse quadro, as redes sociais parecem oferecer um potencial considerável para a entrega de campanhas de saúde pública, por várias razões. Primeiramente, por que informações cujo meio de circulação é a internet costumam alcançar, dependendo do canal e da capacidade de *marketing* do divulgador, centenas e até milhares de pessoas. Em segundo lugar, as mensagens podem ser entregues por meio de contatos

existentes. Terceiro, ao contrário das intervenções mais utilizadas na *web*, as redes sociais geralmente alcançam altos níveis de engajamento e retenção de usuários. Finalmente, a rede social "exige" que os usuários se envolvam ativamente e gerem conteúdo (mesmo que seja de forma inconsciente), o que pode ser bem mais influente do que os *sites* tradicionais e a publicidade que são tipicamente de natureza mais passiva.

Interessados nessas iniciativas, em especial a campanha de prevenção ao câncer de próstata, intitulada Novembro Azul, que teve origem em 2003, quando dois amigos australianos brincavam sobre trazer o bigode de volta, hábito que estava bem fora de moda. Então, inspirados pela mãe de um amigo que levantava fundos para o combate ao câncer de mama (outubro rosa), resolveram fazer o mesmo para o câncer de próstata. A regra: deixar crescer um bigode e cobrar 10 dólares de cada bigodudo. Naquele ano, 30 amigos se reuniram nessa causa (TONIN, 2017).

Atualmente, mais de 20 países aderiram à campanha. No Brasil, a Novembro Azul iniciou em 2008, trazida pelo Instituto Lado a Lado pela Vida, juntamente com a Sociedade Brasileira de Urologia (SBU). Entre as ações promovidas durante o mês, há a oferta de exames de próstata gratuitos ou com desconto, além da promoção de ações que levam informação às pessoas e profissionais. As iniciativas mundiais carregam sempre o termo *Movember*, termo que veio da junção da palavra inglesa *moustache* (bigode) com *november* (novembro), reforçando a importância da campanha. Vale lembrar, também, que o mês foi escolhido por conta do Dia Mundial do Combate ao Câncer de Próstata, comemorado em 17 de novembro (MINUTO SAUDÁVEL, 2018).

A campanha mobiliza pessoas, setores e diversos órgãos no Brasil. Por exemplo, celebridades, postos de saúde, hospitais, propagandas de rádio e televisão. Como já era esperada, a campanha também foi realizada na internet, principalmente nas redes sociais (LADO A LADO PELA VIDA, 2019).

A Figura 1 ilustra a propaganda e o *slogan* da campanha nas redes sociais do ano de 2018.

Figura 1 – Propaganda e o *slogan* da campanha

Fonte: Lado a Lado pela Vida (2019)

É a partir dessa importante iniciativa na área da saúde que este trabalho está inserido, no sentido de trazer contribuições para a prevenção do câncer de próstata.

Por tratar-se de pesquisa que busca dados em rede social sobre manifestações de usuários acerca de campanha de prevenção à doença, torna-se importante definir quais conceitos serão mobilizados. Como o de rede social, de mídia social, de mineração de dados, de análise de sentimentos e de Processamento de Linguagem Natural (PLN).

2.2 REDES SOCIAIS E MÍDIAS SOCIAIS

Uma mídia social é definida como uma estrutura social de indivíduos, que estão relacionados (direta ou indiretamente uns com os outros) com base em uma relação comum de interesse, por exemplo, de amizade, de confiança etc. (INJADAT; SALO; NASSIF, 2017). Já a rede social, segundo o *site* Conceitos (2019), é definida como uma aplicação da *web* cuja finalidade é relacionar as pessoas. Assim, as pessoas que integram uma rede social podem conectar-se entre si e criar vínculos. Elas permitem a criação de um perfil com limitações em sua acessibilidade que pode ser compartilhada ou não com quem solicite. Logo, rede social é um tipo de mídia social.

As redes sociais ocupam crescente espaço no discurso acadêmico, nas mídias e nas organizações. Elas configuram o espaço comunicacional tal qual representado e/ou a experiência do/no mundo globalizado e interconectado no qual se produzem formas

diferenciadas de ações coletivas, de expressão de identidades, conhecimentos, informações e culturas. Segundo Marteleto (2018), indicam mudanças e permanências nos modos de comunicação e transferência de informações, nas formas de sociabilidade, aprendizagem, autorias, escritas e acesso aos patrimônios culturais e de saberes das sociedades.

Portanto, os fenômenos das redes sociais são incontestáveis em todas as esferas da sociedade, as empresas mantêm perfis em *sites*, como Facebook®, Twitter e Google+. Pessoas se relacionam com amigos, partilham opiniões, discordam ou discutem os mais diversos assuntos (GABARDO, 2015).

As redes sociais estão mudando o mundo. Com o advento dos *smartphones* e das redes sociais, a acessibilidade das informações é maior do que nunca. Frequentemente, pede-se aos clientes que "curtam" empresas no Facebook®, "sigam" empresas no Twitter ou "conectem-se" via LinkedIn. Como resultado, os clientes estão cada vez mais conectados às empresas, mais informados sobre as seleções de produtos e mais influentes na relação comprador-vendedor (AGNIHOTRI *et al*., 2016).

O uso das redes sociais (por exemplo, Facebook®, LinkedIn, Twitter) cresceu significativamente entre os consumidores. O Brasil é o segundo país do mundo que passa mais tempo conectado à internet. A média do brasileiro é de 9h29min todos os dias. "Isso que dizer que, dos 365 dias do ano, em 145 deles fica-se conectado à internet. A pesquisa coloca o Brasil apenas atrás das Filipinas, que passa mais de 10 horas por dia conectada à internet." (DA SILVA, 2019). Esses números motivam os pesquisadores a entender como usar as redes sociais para influenciar as preferências do consumidor, as decisões de compra e o valor do produto (MICHAELIDOU; SIAMAGKA; CHRISTODOULIDES, 2011; KUMAR; MIRCHANDANI, 2012).

Além disso, pesquisadores também estudaram o uso das redes sociais para se comunicar com os clientes e explorar a melhora da sua experiência (WILSON *et al*., 2011), e os blocos de construção funcionais de uma estratégia de rede social para atingir os consumidores (KIETZMANN *et al*., 2011).

Existem dois efeitos principais das redes sociais dentro da relação empresa-cliente. Primeiro, a rede social fornece um meio de se comunicar com os clientes de uma maneira que pode plausivelmente permitir uma maior capacidade de resposta do vendedor. Diz-se provavelmente por que, por exemplo, quando reclamações de consumidores são apresentadas em um *site* de rede social, 58% dos consumidores querem uma resposta, no entanto, apenas 22% relatam recebê-la (RIGHTNOW TECHNOLOGIES, 2010).

2.3 MINERAÇÃO DE DADOS E TEXTOS

A internet é de longe, o maior banco de dados mundial, contendo uma imensa quantidade de dados de vários tipos que podem ser consumidos pelos usuários para diversas necessidades. Podem conter dados de valor inestimável para as empresas, se utilizados de forma eficaz. A mineração de dados na *web* é um processo que visa a encontrar informações ou conhecimentos úteis a partir do conteúdo de determinada página da *web* (JAYAMALINI; PONNAVAIKKO, 2017).

Normalmente os dados das redes sociais não são coletados com o intuito de pesquisar, logo, faz-se necessário mudar a estrutura desses dados, pois 80% dos textos disponíveis não estão estruturados, isto é, não possuem uma organização para serem trabalhados no futuro, enquanto apenas 20% estão estruturados (SALLOUM *et al*., 2017).

A maioria das teorias e métodos de mineração de dados é desenvolvida e ampliada com base na teoria estatística. A inteligência artificial (IA) é usada para gerar o processo do pensamento humano, que permite ao computador a capacidade de aprender sem programação precisa e facilita novas técnicas utilizadas nos processos de mineração de dados.

Tecnicamente, a mineração de dados é o processo de encontrar correlações ou padrões entre dezenas de campos em bancos de dados. Ela é usada no cotidiano de vários setores das empresas, no varejo, nas finanças, na saúde, na comunicação e no *marketing*. Ela permite que esses setores determinem as relações entre fatores internos, como preço, posicionamento do produto ou habilidades da equipe e até fatores externos, como, os indicadores econômicos, os seus concorrentes e os dados demográficos de clientes (RAORANE; KULKARNI; 2011). As funcionalidades da mineração de dados são classificadas em: 1) funções descritivas, que visam principalmente a explorar as regras, características e relações potenciais ou recessivas que existem nos dados, como a generalização, associação, mineração de padrões de sequência e agrupamento; e 2) funções preditivas, que geralmente analisam as tendências relevantes dos dados ou as leis relevantes para prever o estado futuro. (CHENG *et al*., 2018). Como exemplo de aprendizado descritivo, pode-se citar a associação entre eventos demandados por pacientes, que podem indicar possíveis relações de causa-efeito ao ser complementada com o respectivo espaço de tempo entre esses eventos associados, podendo ser monitorados em tempo real. Já a função preditiva pode ser, por exemplo, aplicada a uma empresa de cursos *online*, e analisando as informações dos seus alunos, detectando quantos

alunos cancelaram o curso ou ficaram mais de 30 dias sem acessar o ambiente *online*. Ou seja, detectando que esses alunos tem maior chance de cancelar o curso.

CHENG *et al*. (2018) afirmam que o processo geral de mineração de dados inclui esclarecimento de problemas, coleta de dados, pré-processamento, mineração de dados no sentido estrito e interpretação e avaliação de resultados. Considerando que, mineração de dados no sentido estrito refere-se apenas a um passo para gerar um padrão específico usando um algoritmo particular dentro de um limite de eficiência computacional aceitável.

Diante disso, diversas técnicas foram desenvolvidas com propósito de recuperar informações importantes contidas em bases de dados, dando origem à área chamada mineração de textos, que deriva das técnicas de mineração de dados, uma vez que as duas procuram extrair informações úteis em dados não estruturados, sendo que esses tipos de dados criam dificuldades para o seu tratamento (FELDMAN *et al*., 2007).

A principal diferença entre mineração de textos e mineração de dados é que a primeira permite a extração de informações relevantes a partir de texto em linguagem natural não estruturada, já o segundo baseia-se em previsões quantitativas, com base em dados estruturados (que podem ser texto ou não) (SILVA, 2010). Para tornar possível a extração de informação útil a partir de bases de dados de texto não estruturadas, passaram a integrar-se nas ferramentas de mineração de dados, técnicas como a recuperação de informação e sistemas de classificação de termos para análise de linguagem natural (BUTLER ANALYTICS, 2014).

Além dos métodos aplicados na mineração de dados que também são usados na mineração de textos, diversas áreas de pesquisa são igualmente importantes para a extração de conhecimento em bases de textos, por exemplo: Inteligência Computacional (IC), Aprendizagem de Máquina (AM), Recuperação da Informação (RI), Ciência Cognitiva e, muito importante, o PLN, que busca descobrir como os computadores podem ser usados para entender a linguagem humana (DE BRITO, 2017).

2.4 PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

O PLN é uma área da ciência da computação que trabalha com o processamento de linguagem humana escrita ou falada. Pesquisas de PLN têm atraído muitas atenções desde a popularização da internet, pois existem milhões de páginas sendo criadas todos os dias. O sucesso das empresas utilizando a PLN também motiva as pesquisas. Um dos exemplos é a Google, com seu primeiro produto amplamente utilizado, um módulo de recuperação de

informações como parte do seu famoso mecanismo de pesquisa, o Google Busca (PURWARIANTI *et al*., 2016).

O PLN é definido como a habilidade de um computador em processar a mesma linguagem que os humanos usam no seu cotidiano. Não é o objetivo no PLN descrever e, sim, criar soluções para problemas pontuais, principalmente os relacionados com o reconhecimento e a reprodução da linguagem humana. As soluções do PLN normalmente são pensadas em termos de menor custo e maior benefício (ROSA, 2011).

Conforme destacado por Jurafsky e Martin (2009), existem seis níveis de linguagem processadas pelas técnicas de PLN: 1) fonético e fonológico, que é associado à articulação de fonemas e uso em linguagem; 2) morfológico, relacionado às classes gramaticais, por exemplo, artigos, substantivos, verbos, etc., e também a formação de palavras, por exemplo, por raízes, prefixos, sufixos, etc.; 3) nível sintático, que busca encontrar dependências sintáticas entre frases em uma sentença, determinando funções sintáticas, como sujeito e predicado, e na construção de árvores sintáticas que determinam a relação entre os constituintes de uma sentença; (4) nível semântico, que possui um amplo conjunto de análises, como desambiguação de palavras, das chamadas polissêmicas, ou seja, com mais de um significado; 5) pragmática, um nível preocupado com a informação implícita no texto; e por fim; 6) retórica, relacionado à estrutura do texto como um todo e às estratégias utilizadas para organização e criação de texto.

Para Cerqueira *et al*. (2010), as abordagens atuais da PLN se dividem basicamente em quatro categorias principais: simbólica, estatística, conexão e a híbrida. A abordagem simbólica analisa os fenômenos linguísticos e seus paradigmas por meio de regras conhecidas da linguagem. A estatística usa cálculos matemáticos para gerar modelos e regras a partir de exemplos de textos e sentenças. A de conexão é parecida com a abordagem estatística, pois também desenvolve modelos genéricos, entretanto, ela faz uma combinação entre o aprendizado estatístico e outras teorias de representação de conhecimento. Já a denominada híbrida mescla métodos vindos de diferentes abordagens.

Independentemente de já existirem vários métodos fortemente precisos para análise e extração do conhecimento relevante baseado em dados estruturados (por exemplo, tabelas ou banco de dados), a missão de extrair informações úteis em bases de dados não estruturados (texto, discurso etc.), como é o caso das redes sociais, ainda é um importante desafio (O'CONNOR *et al*., 2010; LIU; ZHANG, 2012). A saída para essa questão vem sendo buscada por muitos pesquisadores no subcampo do PLN denominado análise de sentimentos (FORTUNY *et al*., 2012).

2.5 ANÁLISE DE SENTIMENTOS/MINERAÇÃO DE OPINIÃO

É uma técnica automatizada de descoberta de conhecimento que procura padrões escondidos em uma enorme quantidade de textos, como nas redes sociais (MOSTAFA, 2013). Essa ação busca criar e organizar uma base de conhecimento, coletando opiniões (positivas, negativas e neutras) de uma forma mais estruturada e explícita que expressem essas opiniões, críticas e avaliações dos usuários relacionados a temas do cotidiano (FORTUNY *et al*., 2012; SOBKOWICZ; KASCHESKY; BOUCHARD, 2012).

A análise de sentimentos, também chamada de mineração de opinião, é uma área de pesquisa muito estudada atualmente. O seu objetivo principal é pesquisar sobre tópicos, produtos, indivíduos, organizações e serviços, buscando analisar os sentimentos, as opiniões, e as emoções das pessoas (SERRANO-GUERRERO *et al*., 2015). Entende-se que emoção é o resultado de um ou mais sentimentos, sendo que os sentimentos são expressos pelo sujeito por meio de sua opinião (SILVA, 2017).

Assim, a análise de sentimentos comumente trabalha com o reconhecimento da polaridade, ao invés de detectar sentimentos distintos definidos como alegria e tristeza, por exemplo, ou seja, define se um texto é objetivo ou subjetivo, e se um texto subjetivo apresenta sentimentos positivos ou negativos, (BAE; LEE, 2012). Mostafa (2013) conceitua, também, a opinião neutra dentro da polaridade, ao admitir que a análise de sentimentos é um processo que tem como objetivo determinar se a polaridade de um *corpus* textual (documento, frase, parágrafo etc.) tende a ser positivo, negativo ou neutro.

Quanto ao modo de classificar os textos, as abordagens em análise de sentimentos baseiam-se na orientação semântica (aprendizagem não supervisionada) ou na aprendizagem supervisionada, do inglês supervised learning, (SL), e ambas são utilizadas por meio de algoritmos e/ou *softwares* específicos (KONTOPOULOS *et al*., 2013). A abordagem baseada em orientação semântica está centrada em listas predeterminadas de palavras positivas e negativas e, nesse caso, o reconhecimento automatizado da polaridade do texto necessita da frequência dos diferentes tipos de palavras que aparecem no texto (YU; DUAN; CAO, 2013).

A oportunidade de captar automaticamente os sentimentos do público em geral sobre eventos sociais, movimentos políticos, campanhas, preferências de produtos despertou o interesse do mundo dos negócios e da comunidade científica, principalmente pelos empolgantes desafios abertos, e também pelas notáveis repercussões e previsões de mercado (CAMBRIA, 2016).

Importante ressaltar que a análise de sentimentos opera com as dificuldades de classificação de dados, pois palavras não têm somente valor em si, ou seja, diferentes contextos podem atribuir diferentes valores a uma palavra, por exemplo, o verbo “morrer”: [morro de medo dele – negativo; ela morre de amores por ele – positivo.]. Em razão dessa complexidade, é necessário lançar mão de cálculo de polaridade, que classifica textos de acordo com a seu aspecto negativo, positivo ou neutro a partir da classificação de cada palavra semântica (isto é, com sentido) presente na frase. Assim, mesmo que uma frase não denote explicitamente um sentimento, como na frase "criança com suspeita de febre amarela morre no hospital", que apenas, superficialmente, narra um fato, ainda poderá ser classificada como positiva ou negativa para a área da saúde (DE BRITO, 2017).

A análise de sentimentos reúne muitas tarefas, entre elas, a extração e classificação dos sentimentos, a classificação da subjetividade, o resumo de opinião ou detecção de *spam*. Para realizar qualquer uma dessas atividades, a análise de sentimentos precisa enfrentar grandes desafios.

A aprendizagem supervisionada refere-se à elaboração de um classificador de sentimentos baseado em um grupo de treinamento. Um grupo de treinamento é uma amostra de dados usada para o treinamento manual de um classificador automático, isto é, um conjunto de dados submetido à classificação humana para que o computador via recursos de inteligência artificial, cria padrões sobre os atributos diversificados de documentos (VINODHINI; CHANDRASEKARAN, 2012).

Ainda sobre a aprendizagem supervisionada, ela apresenta bons resultados quanto à precisão da classificação automatizada dos dados. No momento atual, a análise de sentimentos está recebendo muita atenção por causa da variedade de suas aplicações diretas, como análises de produtos, serviços, perfil do público, tendências políticas, dentre outras. Inclusive questões sobre qualidade de vida e segurança pública podem ser descobertas, monitoradas e suavizadas por meio da análise de informação das redes sociais, reconhecendo padrões e tendências significativas (KAVANAUGH *et al*., 2012).

O fato de muitas pessoas se expressarem nas redes sociais sobre qualquer assunto tornam as opiniões menos parciais e, consequentemente, mais sinceras. Cáceres (2011) acredita que os usuários das redes sociais têm muito a dizer, muito a ensinar e muitas tendências a propor. Como resultado, as opiniões demonstradas nessas mídias são cada vez mais consideradas no processo de tomada de decisão e na obtenção de um retorno imparcial sobre vários assuntos. Com efeito, o trabalho com diversas informações das redes sociais necessitará de grandes investimentos (MONTOYO; MARTÍNEZ-BARCO; BALAHUR, 2012).

3 METODOLOGIA

3.1 NATUREZA DO ESTUDO

O presente estudo é de natureza quantitativa, pois desenvolve, analisa um modelo classificador automático de polaridade de mensagens. As postagens submetidas ao modelo são as de seguidores da página denominada Novembro Azul no Facebook®, em época de campanha de prevenção do câncer de próstata. É testada a acurácia do modelo com métodos da linguagem de programação Python. Se bem-sucedido no seu desenvolvimento e nos testes, o modelo poderá ser utilizado como ferramenta auxiliadora nos processos de tomada de decisão dos gestores e profissionais de saúde de campanhas de prevenção de doenças.

3.2 TIPO DE ESTUDO

Pode-se considerar o estudo como uma produção tecnológica inovadora e estudo quantitativo, pois além de buscar e utilizar ferramentas e/ou métodos de análise de sentimentos o estudo visa a preencher uma lacuna, visto que trabalhou com uma rede social que parece pouco explorada para fins de pesquisa e, principalmente, por se dedicar a um tema pouco investigado em pesquisas na área da saúde.

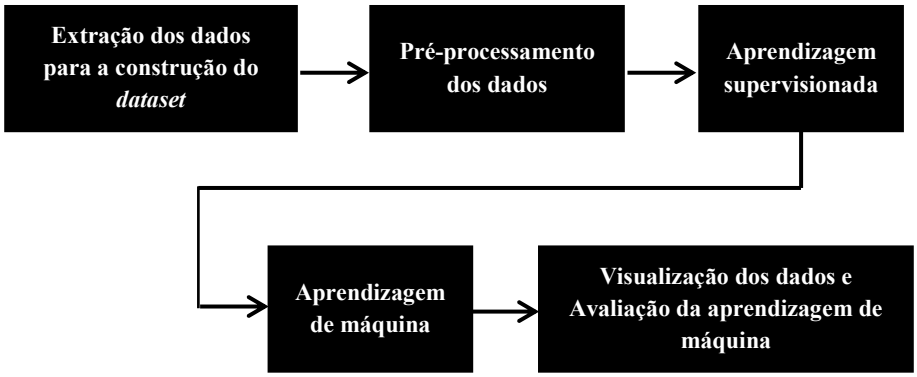
3.3 PROTOCOLO DO ESTUDO

O estudo foi desenvolvido em cinco etapas, baseadas em De Brito (2017) e França e Oliveira (2013), listadas a seguir e detalhadas individualmente nas próximas seções:

1. Extração dos dados para a construção do *dataset*.
2. Pré-processamento dos dados.
3. Aprendizagem supervisionada (treinamento).
4. Aprendizagem de máquina.
5. Visualização dos dados e avaliação da aprendizagem de máquina.

A Figura 2 ilustra as etapas do estudo.

Figura 2 – Etapas do estudo



Fonte: Elaborada pelo autor

Na primeira etapa do estudo, que representa a coleta de dados, foi construído um *dataset,* ou seja, foi criada uma planilha como base de dados, com 645 postagens, originada pela busca e extração de postagens e de comentários relacionados ao câncer de próstata, de 1º a 30 de novembro de 2018, mês da campanha, postados na *fanpage* denominada Novembro Azul, no Facebook®, essa *fanpage* é voltada somente para a campanha de prevenção ao câncer de próstata,, de modo que o modelo a ser implementado fosse capaz de realizar classificações.

A seguir, na segunda etapa, realizou-se o pré-processamento dos dados, no qual foram removidas as palavras que não tinham importância para o processo de análise de sentimentos de modo que os textos pudessem ser estruturados em uma representação. Ainda nessa etapa, foram extraídos os sufixos das palavras, pois no processo de análise de sentimentos se faz necessário ter somente o radical da palavra. Foi realizada também a exclusão de linhas em branco da base, e foram deletadas as palavras que se repetiam, resultando em um total de 261 postagens, ou seja, 384 mensagens/linhas foram descartadas, pois consideradas inválidas. Isto é, essa etapa teve por objetivo melhorar a qualidade e a organização dos dados disponíveis, os quais foram submetidos a um algoritmo de mineração de texto.

Na terceira etapa, foi realizada a classificação dos dados, na qual foram categorizadas opiniões inferidas das postagens. Essa classificação foi realizada em uma base de dados com 182 mensagens, ou seja, 70% do total das publicações válidas, chamada de treinamento,

executada por classificadores humanos. Nesse caso, por três especialistas da área de tecnologia.

A quarta etapa foi voltada à utilização e ao incremento do algoritmo de aprendizagem de máquina, denominado *naive bayes*. Nessa etapa, o algoritmo analisou, processou e treinou com o material que os especialistas (pessoas) utilizaram na etapa anterior, classificando de forma automática as 79 mensagens restantes na base de dados, ou seja, os outros 30% das mensagens válidas.

Na quinta e última etapa, os resultados foram visualizados e o algoritmo *naive bayes* avaliado com o teste de acurácia, a saber, a porcentagem de seus acertos e erros.

Vale a pena ressaltar que durante a preparação do código e alguns testes do código utilizou-se uma base de dados, chamada no decorrer do estudo de base aleatória (Apêndice A), e que não é a da página Novembro Azul. Esse tipo de ação é utilizada com frequência, pois o objetivo dessa abordagem é a de não deixar o código com o chamado “ciclo vicioso”, pois o modelo “acabará se acostumando” com cada frase, e terá dificuldade de classificar frases que ele não treinou. A suposta facilidade de classificar frases treinadas não contribui para a o sucesso do trabalho.

4 ASPECTOS ÉTICOS DO ESTUDO

Este estudo está em conformidade com a Resolução n° 466/2012, do Conselho Nacional de Saúde (CNS), que oferta os termos e condições a serem seguidos em todas as pesquisas que envolvam humanos. O documento trata das exigências do sistema de avaliação ética brasileiro, que busca a proteção da integridade dos participantes de pesquisa (MINISTÉRIO DA SAÚDE, 2012).

Segundo a resolução do CNS, as pesquisas envolvendo seres humanos devem atender às exigências éticas e científicas fundamentais, a eticidade de uma pesquisa acarreta em:

> a) Consentimento Livre e Esclarecido dos indivíduos-alvo e a proteção a grupos vulneráveis e aos legalmente incapazes (autonomia). Nesse sentido, a pesquisa envolvendo seres humanos deverá sempre tratá-los em sua dignidade, respeitá-los em sua autonomia e defendê-los em sua vulnerabilidade;
> b) ponderação entre riscos e benefícios, tanto atuais como potenciais, individuais ou coletivos (beneficência), comprometendo-se com o máximo de benefícios e o mínimo de danos e riscos; c) garantia de que danos previsíveis serão evitados (não maleficência);
> d) relevância social da pesquisa com vantagens significativas para os sujeitos da pesquisa e minimização do ônus para os sujeitos vulneráveis, o que garante a igual consideração dos interesses envolvidos, não perdendo o sentido de sua destinação sócio- humanitária (justiça e equidade);
> e) anonimato, que consiste no zelo das informações confidenciais e dados obtidos (MINISTÉRIO DA SAÚDE, 2012).

A presente pesquisa foi dispensada da apreciação pelo Comitê de Ética em Pesquisa, pois trabalha exclusivamente com a análise de postagens de usuários da rede social pública Facebook®, e não traz identificações de quem as inseriu na rede social, logo, o nome do usuário, a sua localização ou algo que possa de alguma maneira expor suas informações não será exibido. Além disso, a pesquisa trabalha somente com o comentário do usuário, e o próprio Facebook® se resguarda com a sua política de acesso aos dados, seja na exibição ou compartilhamento dos dados desses usuários ou na sua distribuição por meio da sua API. Apoia-se também na Lei n. 13.709, de 14 de agosto de 2018, que dispõe sobre a proteção de dados pessoais, na internet.

5 DESENVOLVIMENTO DO MÉTODO (MODELO) E SEUS RESULTADOS

Nesta seção, apresenta-se como foi implementada cada uma das etapas do protocolo bem como de suas fases de desenvolvimento.

5.1 EXTRAÇÃO DOS DADOS PARA A CONSTRUÇÃO DO *DATASET* – ETAPA 1

A construção do *dataset* consistiu em buscar e extrair as postagens e comentários relacionados ao câncer de próstata, da página denominada Novembro Azul, no Facebook®. Como ferramenta de extração foi utilizada a estrutura denominada Netlytic, do *site* netlytic.org, que atualmente fornece gratuitamente a estudantes e pesquisadores os serviços de análise de redes sociais e tarefas de mineração de texto.

Figura 3 – Tela da plataforma Netlytic

Fonte: Netlytic (2019)

A tecnologia utiliza foram as APIs públicas de *sites* de redes sociais (NETLYTIC, 2019), para fazer a coleta das postagens e comentários. É uma ferramenta de análise de textos e redes sociais que permite aos usuários capturar e importar os dados das conversações

online, podendo visualizar e explorar temas emergentes (GRUZD; PAULIN; HAYTHORNTHWAITE, 2016).

Segundo o Netlytic (2019), a plataforma é ideal para ensinar e aprender sobre análises de redes sociais. Atualmente, está sendo usada por centenas de educadores e milhares de estudantes em todo o mundo. O Netlytic apoia a pesquisa em ciências sociais sobre a participação *online* das mais variadas comunidades e promove pesquisas éticas com dados de mídia social. Isso significa que respeita os termos de serviço dos *sites* e a privacidade dos usuários de redes sociais.

A plataforma utiliza a API do Facebook® para realizar as buscas; segundo Netlytic (2019), existem diferentes critérios e limites para a utilização de cada API no Facebook®, sendo eles:

- necessidade de vincular a conta do Facebook® do pesquisador para usar o importador;
- o importador usa a API (v2.7) de gráficos do Facebook® ;
- retorna mensagens e respostas de páginas públicas do Facebook®;
- retorna até 100 postagens de nível superior de/para uma página bem como até 25 respostas por postagem;
- não inclui respostas a comentários.

A API do Facebook® utilizada na plataforma não faz a busca por palavras-chave, mas, sim, pelo nome da *fanpage* a ser explorada, nesse caso, a denominada Novembro Azul. A API também permite a extração por um período pré-determinado. Neste estudo foi o do mês de novembro de 2018, que é o mês da Campanha Novembro Azul.

A Figura 4 mostra a tela de configurações de critérios de busca na plataforma.

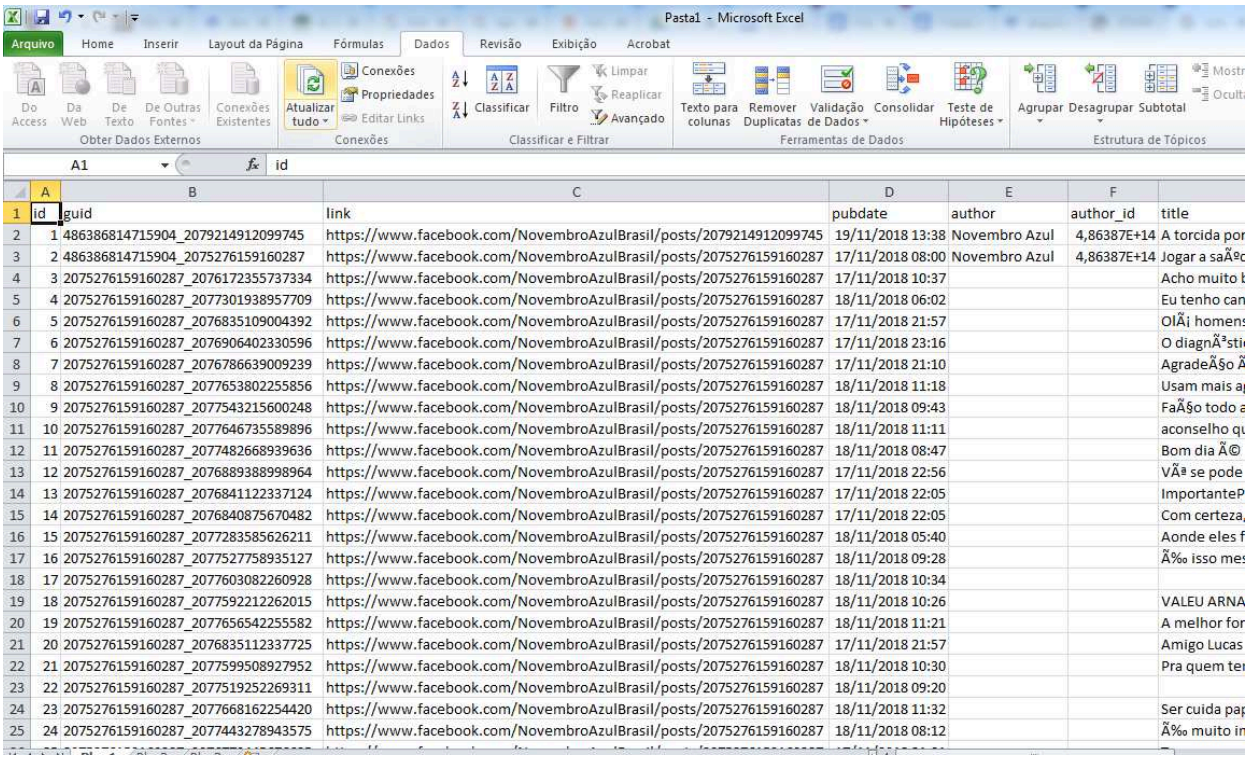
Figura 4 – Tela de configuração da plataforma

Fonte: Netlytic (2019)

Nessa primeira etapa, que foi a extração dos dados e, a partir disso, a criação do *dataset,* obteve-se como resultado um total de 645 postagens e comentários da *fanpage* Novembro Azul.

Em seguida, esses dados foram exportados no formato *Comma-Separated Values* (CSV), que significa valores separados por vírgula; atualmente, esse formato é muito utilizado para pesquisas de mineração de texto. A Figura 5 ilustra parte dos dados que foram exportados.

Figura 5 – Amostra dos dados exportados



	A	B	C	D	E	F	
1	id	guid	link	pubdate	author	author_id	title
2	1	486386814715904_2079214912099745	https://www.facebook.com/NovembroAzulBrasil/posts/2079214912099745	19/11/2018 13:38	Novembro Azul	4,86387E+14	A torcida por
3	2	486386814715904_2075276159160287	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	17/11/2018 08:00	Novembro Azul	4,86387E+14	Jogar a saÃºd
4	3	2075276159160287_2076172355737334	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	17/11/2018 10:37			Acho muito b
5	4	2075276159160287_2077301938957709	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	18/11/2018 06:02			Eu tenho can
6	5	2075276159160287_2076835109004392	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	17/11/2018 21:57			OlÃ¡ homens
7	6	2075276159160287_2076906402330596	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	17/11/2018 23:16			O diagnÃ³sti
8	7	2075276159160287_2076786639009239	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	17/11/2018 21:10			AgradeÃ§o Ã
9	8	2075276159160287_2077653802255856	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	18/11/2018 11:18			Usam mais ap
10	9	2075276159160287_2077543215600248	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	18/11/2018 09:43			FaÃ§o todo a
11	10	2075276159160287_2077646735589896	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	18/11/2018 11:11			aconselho qu
12	11	2075276159160287_2077482668939636	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	18/11/2018 08:47			Bom dia Ã©
13	12	2075276159160287_2076889388998964	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	17/11/2018 22:56			VÃª se pode
14	13	2075276159160287_2076841122337124	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	17/11/2018 22:05			ImportanteP
15	14	2075276159160287_2076840875670482	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	17/11/2018 22:05			Com certeza,
16	15	2075276159160287_2077283585626211	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	18/11/2018 05:40			Aonde eles f
17	16	2075276159160287_2077527758935127	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	18/11/2018 09:28			Ã‰ isso mes
18	17	2075276159160287_2077603082260928	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	18/11/2018 10:34			
19	18	2075276159160287_2077592212262015	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	18/11/2018 10:26			VALEU ARNA
20	19	2075276159160287_2077656542255582	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	18/11/2018 11:21			A melhor for
21	20	2075276159160287_2076835112337725	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	17/11/2018 21:57			Amigo Lucas
22	21	2075276159160287_2077599508927952	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	18/11/2018 10:30			Pra quem te
23	22	2075276159160287_2077519252269311	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	18/11/2018 09:20			
24	23	2075276159160287_2077668162254420	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	18/11/2018 11:32			Ser cuida pa
25	24	2075276159160287_2077443278943575	https://www.facebook.com/NovembroAzulBrasil/posts/2075276159160287	18/11/2018 08:12			Ã‰ muito im

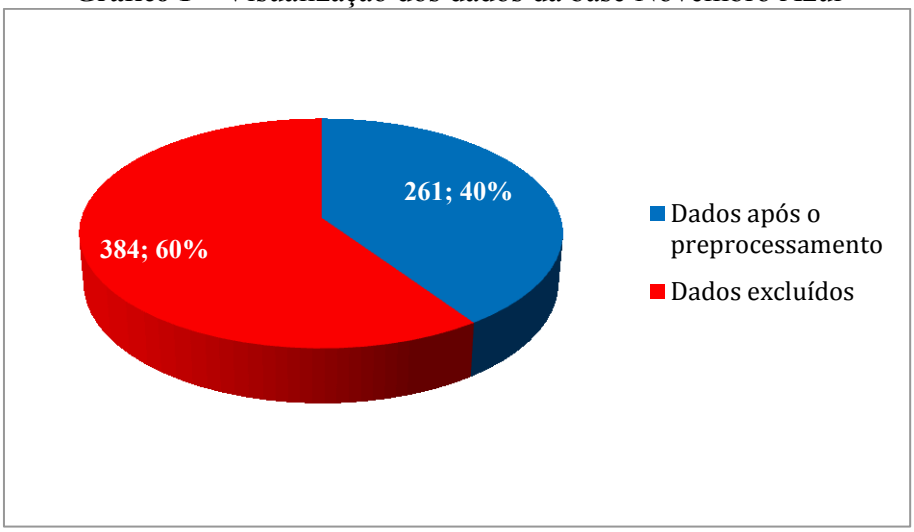
Fonte: Elaborada pelo autor a partir de Netlytic (2019)

5.2 PRÉ-PROCESSAMENTO DOS DADOS – ETAPA 2

Nessa etapa, chamada de pré-processamento, os dados passam por algumas filtragens. O primeiro filtro aplicado foi o de exclusão de linhas em branco, frases repetidas, frases neutras, exclusão de caracteres especiais, de abreviações e das colunas com categorias irrelevantes para a pesquisa, deixando somente a coluna com a postagem ou comentário do usuário. É nessa coluna onde estão os dados que interessam para a pesquisa. Nessa etapa restaram 261 mensagens.

O que isso significa em termos de porcentagem está demonstrado no Gráfico 1.

Gráfico 1 – Visualização dos dados da base Novembro Azul



Fonte: Elaborado pelo autor

Em seguida, as mensagens passaram por um algoritmo de remoção das denominadas *stopwords*. As *stopwords* são palavras sem significados, que não agregam na frase, como, artigos, preposições, advérbios, ou seja, as chamadas palavras gramaticais, por exemplo: a, o, que, do, da, para, com, uma, se, como, mas, isso, depois etc. Para essa função é indicado trabalhar com a linguagem de programação chamada Python e utilizar a sua biblioteca com recursos de PNL, denominada Natural Language Tool Kit (NLTK – Processamento de Linguagem Natural). Nesse acervo, palavras são categorizadas como *stopwords* do português e, caso alguma delas esteja presente na mensagem sendo analisada, é retirada do texto.

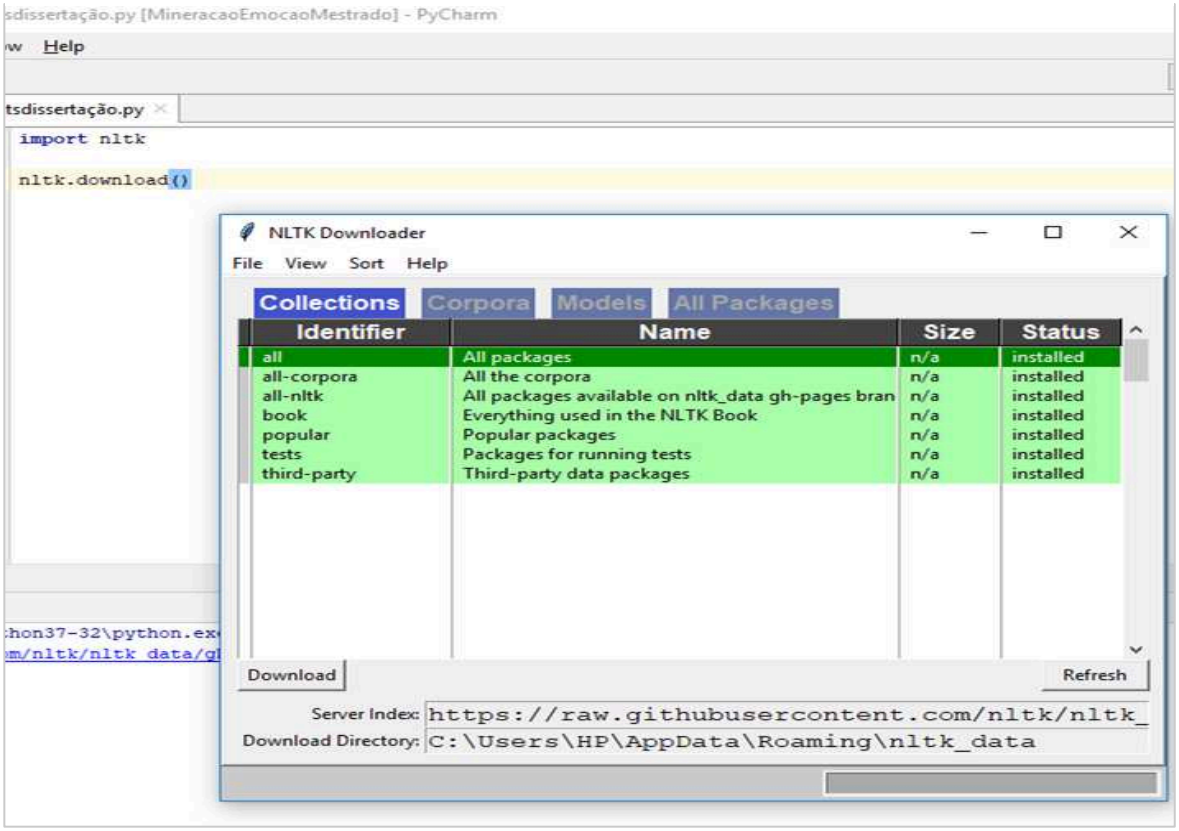
Foi aplicado também um método chamado FreqDist, que calcula a frequência das palavras e logo depois é aplicado um filtro para deixá-las únicas, isto é, remover as palavras repetidas, tudo isso usando também a biblioteca NLTK.

O último filtro dessa etapa é a aplicação de outro método da biblioteca NLTK denominado de *stemming*, que busca manter na mensagem somente os radicais das palavras. Por exemplo, o verbo ENCONTRAR, aplicando a técnica de *stemming* resultaria em ENCONT, que como podemos observar removeu o sufixo da palavra (DE FRANÇA; OLIVEIRA, 2013). Esse método identifica similaridades em função da morfologia das palavras, reduzindo o número de atributos do texto, visto que palavras com morfologia semelhante representam de forma genérica o mesmo conceito (ALVAREZ, 2014).

Instalado o Python e o ambiente de desenvolvimento, foi criado um projeto no qual deveria conter todos os arquivos Python utilizados no trabalho. O primeiro processo foi

importar a biblioteca NLTK, e baixar todos os pacotes dessa biblioteca, que é responsável por trabalhar com o PLN, essencial para o trabalho proposto.

Figura 6 – Importação da biblioteca NLTK e download de seus pacotes



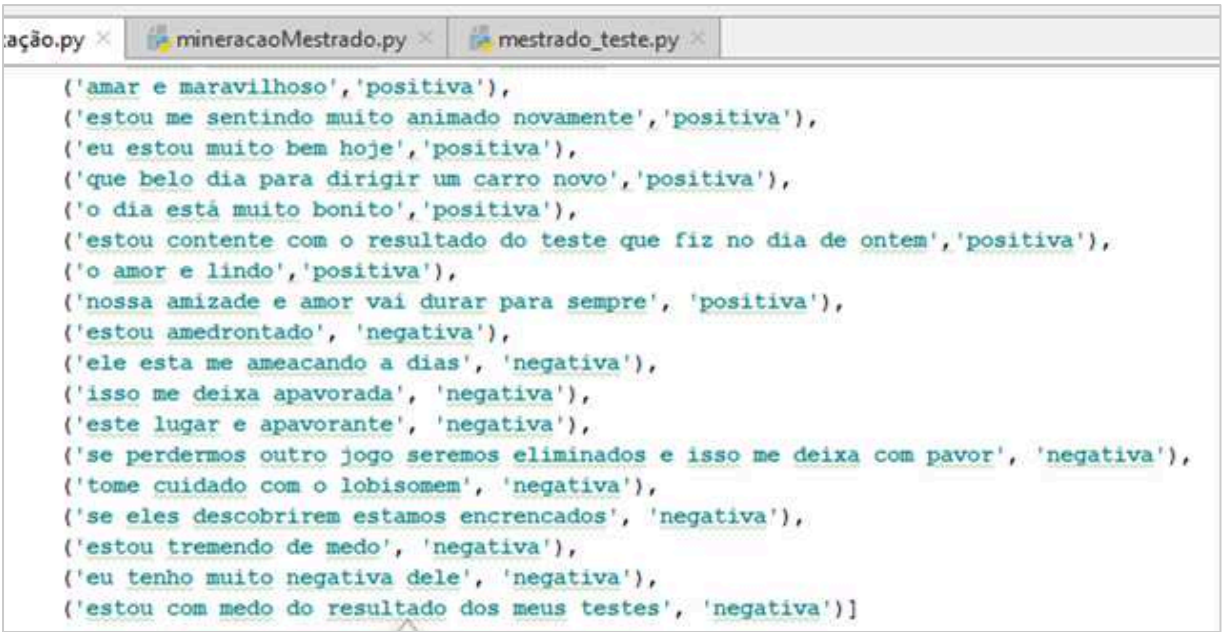
Fonte: Elaborada pelo autor

Logo em seguida, foi criada uma variável, denominada base, que recebeu frases positivas e negativas. Essas frases serviram para realizar testes, sendo que ainda não eram as retiradas do banco de dados.[3] O primeiro teste realizado com as frases foi o de impressão[4] desses textos na tela, como se pode observar na figura a seguir.

[3] Frases retiradas do banco de dados do curso Mineração de Emoção em Textos com Python e NLTK, disponível em: https://www.udemy.com/course/mineracao-de-emocao-em-textos-com-python-e-nltk/learn/lecture/7317124#overview. Acesso em: 27 ago. 2019.

[4] Impressão em linguagem computacional significa mostrar, apresentar.

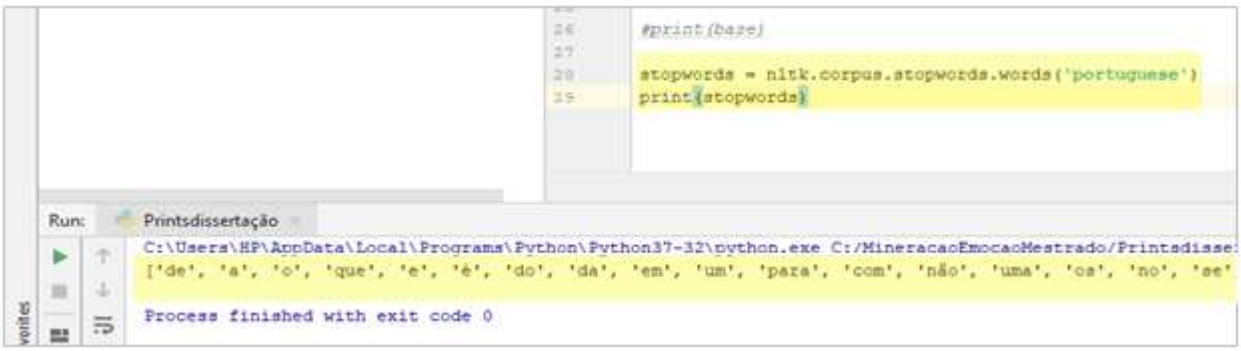
Figura 7 – Frases e suas classificações



Fonte: Elaborada pelo autor

Em seguida, foram removidas as *stopwords*, conforme quadro a seguir. Ou seja, palavras que não influenciavam nos resultados dos algoritmos isto é, não determinavam se a frase iria ser classificada como positiva ou negativa.

Figura 8 – Código que mostra as *stopwords* e exemplos



Fonte: Elaborada pelo autor

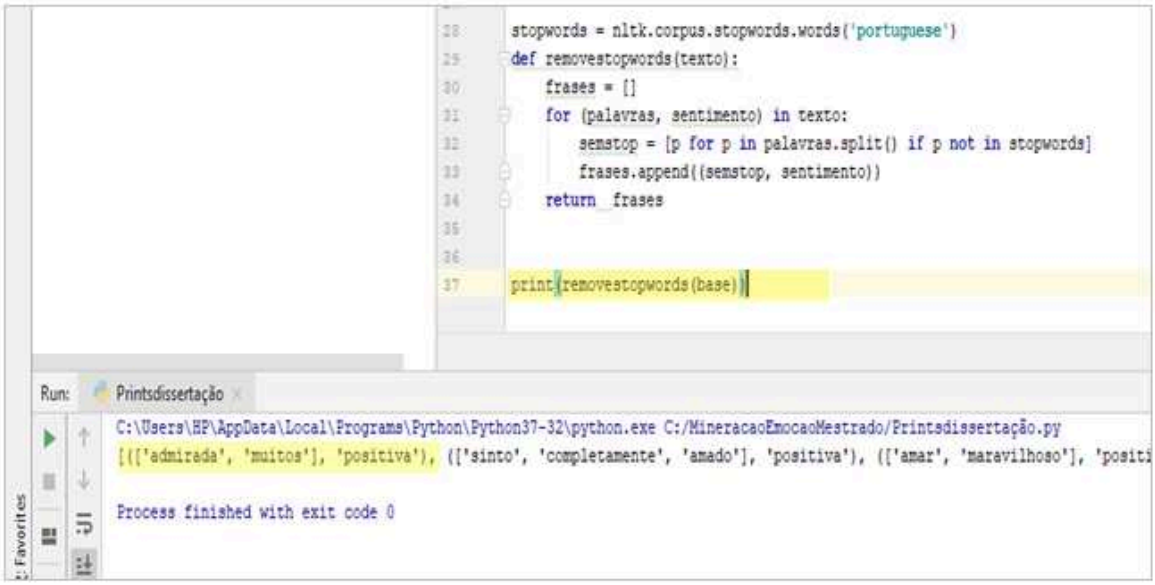
Quadro 1 – Lista de *stopwords* da biblioteca NLTK

'de', 'a', 'o', 'que', 'e', 'é', 'do', 'da', 'em', 'um', 'para', 'com', 'não', 'uma', 'os', 'no', 'se', 'na', 'por', 'mais', 'as', 'dos', 'como', 'mas', 'ao', 'ele', 'das', 'à', 'seu', 'sua', 'ou', 'quando', 'muito', 'nos', 'já', 'eu', 'também', 'só', 'pelo', 'pela', 'até', 'isso', 'ela', 'entre', 'depois', 'sem', 'mesmo', 'aos', 'seus', 'quem', 'nas', 'me', 'esse', 'eles', 'você', 'essa', 'num', 'nem', 'suas', 'meu', 'às', 'minha', 'numa', 'pelos', 'elas', 'qual', 'nós', 'lhe', 'deles', 'essas', 'esses', 'pelas', 'este', 'dele', 'tu', 'te', 'vocês', 'vos', 'lhes', 'meus', 'minhas', 'teu', 'tua', 'teus', 'tuas', 'nosso', 'nossa', 'nossos', 'nossas', 'dela', 'delas', 'esta', 'estes', 'estas', 'aquele', 'aquela', 'aqueles', 'aquelas', 'isto', 'aquilo', 'estou', 'está', 'estamos', 'estão', 'estive', 'esteve', 'estivemos', 'estiveram', 'estava', 'estávamos', 'estavam', 'estivera', 'estivéramos', 'esteja', 'estejamos', 'estejam', 'estivesse', 'estivéssemos', 'estivessem', 'estiver', 'estivermos', 'estiverem', 'hei', 'há', 'havemos', 'hão', 'houve', 'houvemos', 'houveram', 'houvera', 'houvéramos', 'haja', 'hajamos', 'hajam', 'houvesse', 'houvéssemos', 'houvessem', 'houver', 'houvermos', 'houverem', 'houverei', 'houverá', 'houveremos', 'houverão', 'houveria', 'houveríamos', 'houveriam', 'sou', 'somos', 'são', 'era', 'éramos', 'eram', 'fui', 'foi', 'fomos', 'foram', 'fora', 'fôramos', 'seja', 'sejamos', 'sejam', 'fosse', 'fôssemos', 'fossem', 'for', 'formos', 'forem', 'serei', 'será', 'seremos', 'serão', 'seria', 'seríamos', 'seriam', 'tenho', 'tem', 'temos', 'tém', 'tinha', 'tínhamos', 'tinham', 'tive', 'teve', 'tivemos', 'tiveram', 'tivera', 'tivéramos', 'tenha', 'tenhamos', 'tenham', 'tivesse', 'tivéssemos', 'tivessem', 'tiver', 'tivermos', 'tiverem', 'terei', 'terá', 'teremos', 'terão', 'teria', 'teríamos', 'teriam'

Fonte: Elaborado pelo autor

Foi criada uma função que retirava as *stopwords* das frases e mostrava na tela as frases sem essas palavras e com a respectiva categoria da frase. A primeira frase da lista é: Exemplo: eu sou admirada por muitos – categoria positiva. Depois da remoção das *stopwords,* a frase ficou: admirada muitos.

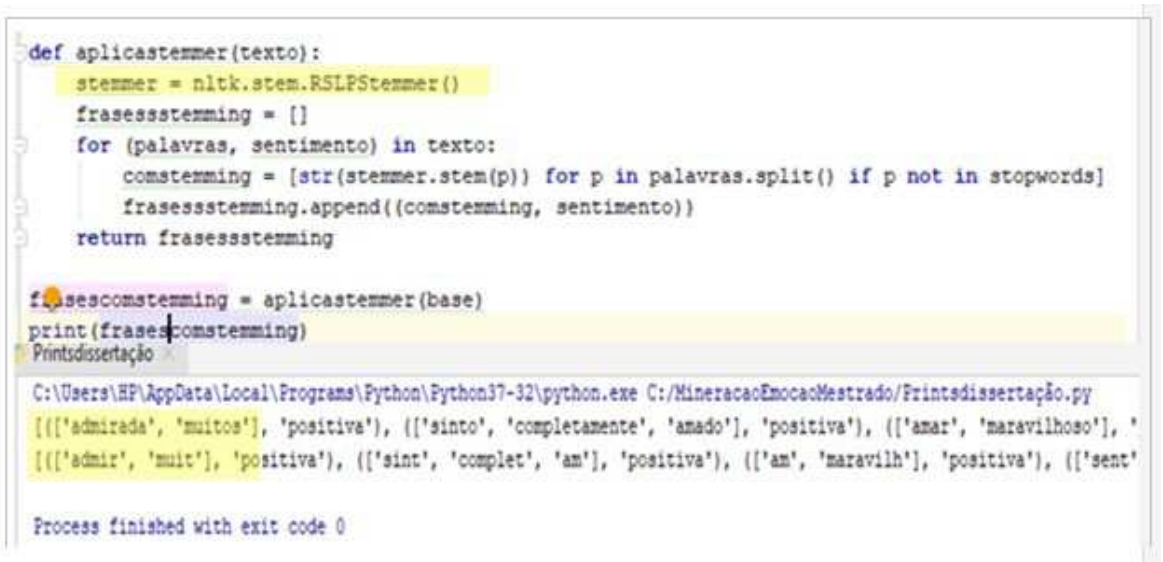
Figura 9 – Código que remove *stopwords* e seu resultado



Fonte: Elaborada pelo autor

Outra etapa importante é a utilização de uma de *stemming*, que deixa na frase somente o radical das palavras. Para essa etapa também foi criada uma função, na qual foi analisada cada palavra. Na biblioteca NLTK, tem-se o stem. RSLOStemmer que faz isso, como se pode ver na figura a seguir.

Figura 10 – Código que exclui prefixos e sufixos e a sua impressão na tela



Fonte: Elaborada pelo autor

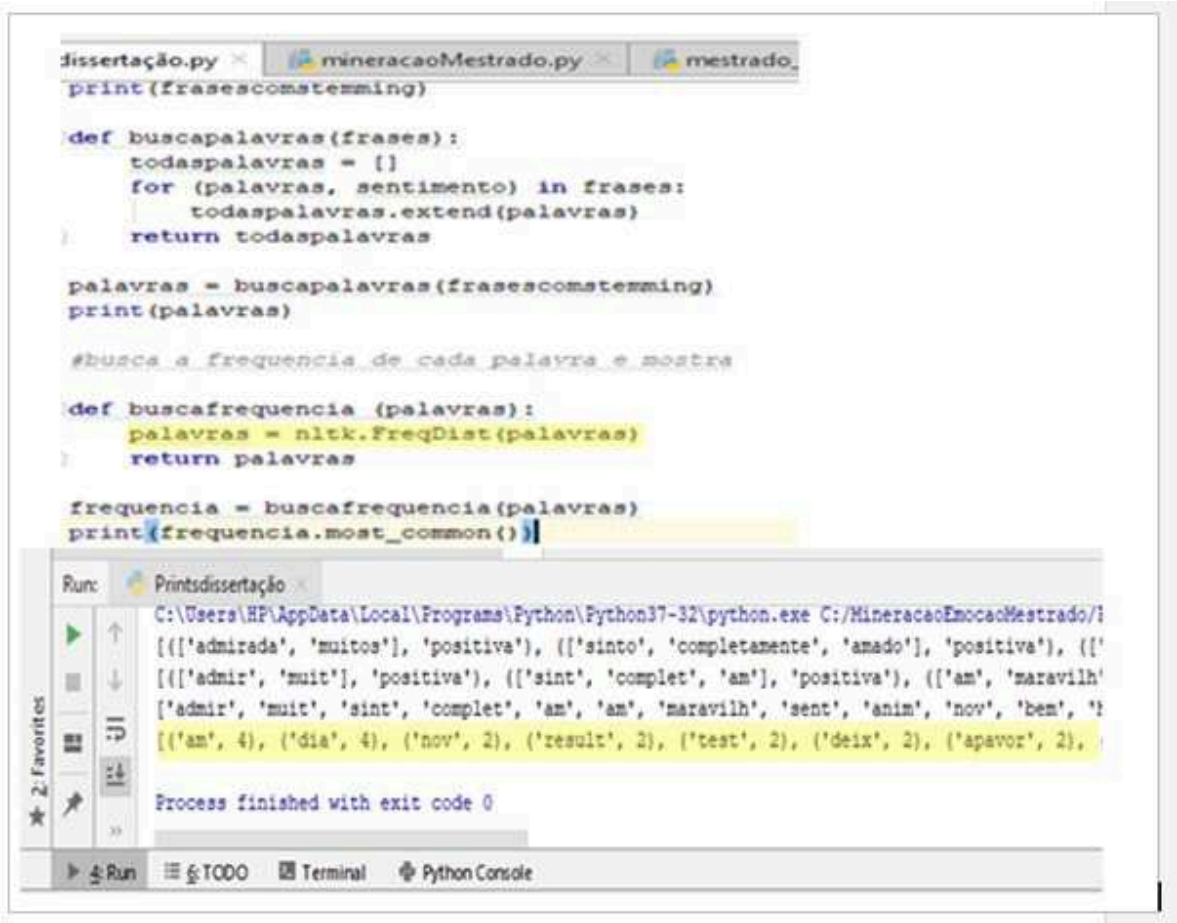
A seguir foi criada uma função que mostra todas as palavras da base de dados sem os sentimentos classificados. Pode-se observar na imagem a seguir todo o incremento do código e também a evolução dos resultados. Primeiro as frases completas, com o seu respectivo sentimento classificado. Logo a seguir o resultado da aplicação do método *stemming*, e, por último, as frases somente com seus radicais, sem a classificação do sentimento.

Figura 11 – Impressão de toda a base e dos radicais

Fonte: Elaborada pelo autor

Outra função importante é a que mostra o número de vezes que cada palavra aparece em todas as frases, pois quanto mais repetir alguma palavra, mais relevância essa palavra terá no treinamento. Nesse caso é utilizado o método da biblioteca NLTK chamado FreqDist. Pode-se observar o código e a sua saída logo a seguir na Figura 12.

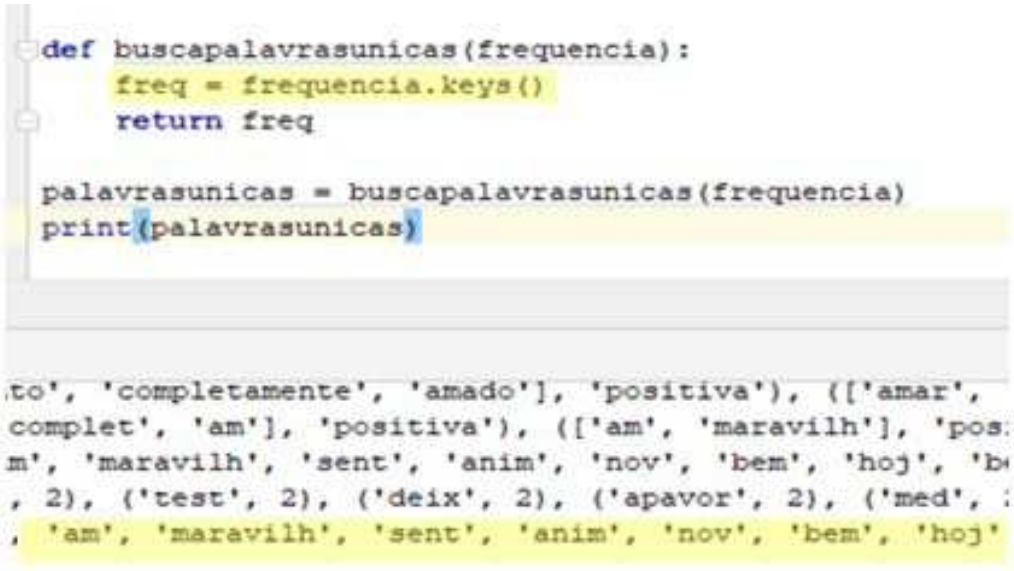
Figura 12 – Frequência das palavras



Fonte: Elaborada pelo autor

Em seguida, foi utilizada uma função que mostra somente as palavras, sem o número de vezes que elas apareciam. O nome do método é frequencia.keys. A seguir também o código e saída.

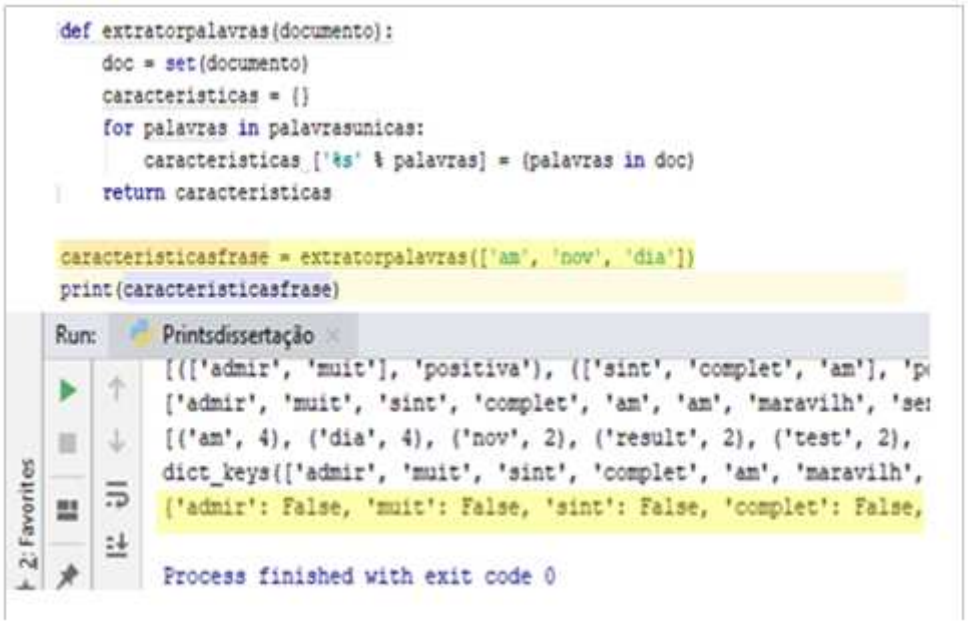
Figura 13 – Código que imprime as palavras da base sem a sua frequência



Fonte: Elaborada pelo autor

Na etapa seguinte, foi criada uma função que recebe algumas palavras e analisa quais dessas palavras estão ou não na base de dados Para essa funcionalidade foi realizado um teste com três palavras. Nesse teste, cada palavra recebida foi comparada com as que já estão na base de dados; se a palavra existe na base é mostrado *True*, se a palavra ainda não está na base é mostrado *False*. Como teste foram inseridas algumas palavras aleatórias. As palavras testadas foram “amor, novo, dia”. O resultado do código é impresso conforme se observa na figura a seguir, apresentando somente os seus radicais.

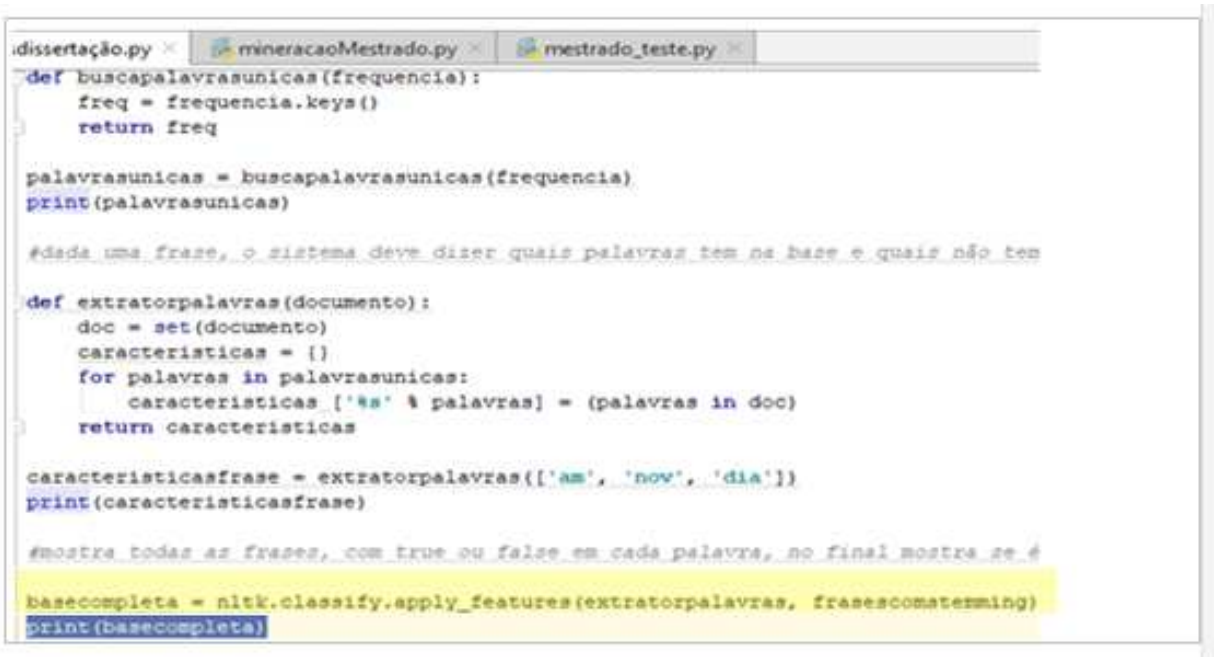
Figura 14 – Código que analisa e insere palavras na base



Fonte: Elaborada pelo autor

A próxima ação foi criar uma função que no final de cada sentença apresentasse a classificação da frase, isto é, se é positiva ou negativa. Para isso foi utilizado o método nltk.classify.apply_features, mostrando assim a característica ou classe que a frase pertence (Figura 15).

Figura 15 – Código da função que mostra a frase e sua classificação ou classe



Fonte: Elaborada pelo autor

5.3 APRENDIZAGEM SUPERVISIONADA – ETAPA 3

Logo depois do pré-processamento dos dados, foi realizada a terceira etapa, chamada de aprendizagem supervisionada, que classifica todos os dados, isto é, todas as 261 mensagens. 182 mensagens (70%) dessa classificação será utilizada na próxima etapa, servindo como uma espécie de treinamento para o algoritmo, já as 79 mensagens (30%) vão servir para o algoritmo classificar automaticamente e depois comparar com a classificação realizada pelos especialistas. Assim, foram analisadas e rotuladas (Apêndice A) por três profissionais da área de tecnologia.

Tabela 1 – Dados da base Novembro Azul

Base de dados	**Nº de frases**	**Porcentagem**
Base treinamento positivas	141	77%
Base treinamento negativas	41	23%
Total frases treinamento	**182**	**70%**
Base teste positivas	61	77%
Base teste negativas	18	23%
Total frases teste	**79**	**30%**
Total de frases	**261**	**100%**

Fonte: Elaborada pelo autor

Esses rótulos são de mensagens consideradas positivas (apoio à Campanha Novembro Azul), negativas (repúdio à Campanha Novembro Azul), ou neutras (que não apoiam e nem repudiam a Campanha, ou seja, não são nem positivas nem negativas), que foram excluídas.

Essa classificação teve como objetivo cruzar os votos de cada pesquisador e decidir qual sentimento classificaria cada mensagem, criando assim uma base de dados chamada de treinamento, que é de suma importância e a grande referência para próxima etapa (aprendizagem de máquina), realizada por algoritmos de aprendizagem de máquina.

5.4 APRENDIZAGEM DE MÁQUINA – ETAPA 4

Na quarta etapa, o foco foi utilizar e incrementar o algoritmo de aprendizagem de máquina, do inglês *machine learning*. Basicamente esse algoritmo checou qual é a probabilidade de um evento acontecer, baseado em eventos anteriores. O mais utilizado para esse tipo de abordagem é o algoritmo *naive bayes*, que é um sistema de classificação que independe de linguagem e que tem apresentado bons resultados na literatura, por exemplo, as ferramentas AntiSpam utilizadas nos *e-mails* (LUCCA *et al*., 2013).

O algoritmo trabalhou com a referência dos dados já classificados por especialistas na fase anterior (treinamento). Esse algoritmo atuou com modelos estatísticos de probabilidades de aprendizagem e fez análises na base de treinamento. Isso gerou um modelo que logo em seguida foi utilizado para testar os outros 30% das mensagens coletadas (base de dados de teste também classificados por humanos), comparando a classificação que o modelo apresentou com a classificação humana, listando assim os erros e os acertos do modelo, medindo a acurácia do sistema de classificação automática.

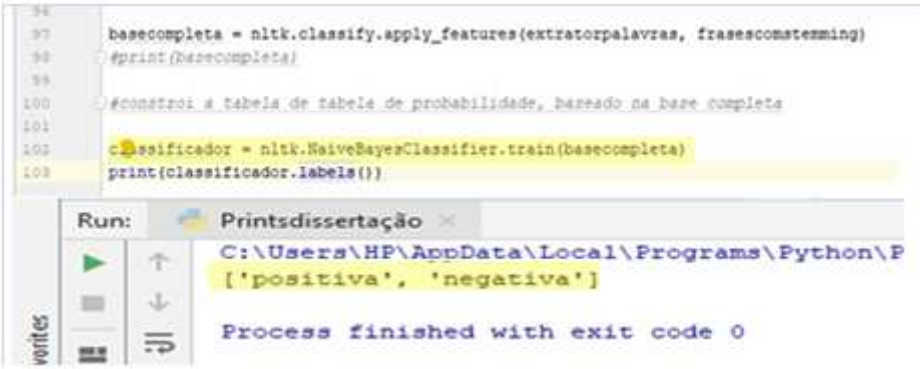
Pode-se observar na Figura 16, que o algoritmo já consegue classificar como positiva ou negativa uma frase, através do método chamado classificador.classify, e também mostra a porcentagem de positiva ou negativa de uma frase, isso através do método classificador.prob_classify. Mas é importante ressaltar que a base de dados que foi utilizada como treinamento é uma base aleatória, ou seja, ainda não são os dados da campanha Novembro Azul.

Figura 16 – Código que classifica e mostra a classe em que a frase pertence

Fonte: Elaborada pelo autor

Em seguida, foi iniciado o treinamento com o algoritmo *naive bayes*, com o método NaiveBayes.train, com referência na base aleatória completa. Além do treinamento realizado, são mostradas as classes das frases na base de dados. Isso por meio do método labels, que disponibiliza na tela as classes que existem na base de dados classificada anteriormente pelos especialistas humanos. Na Figura 17 temos o resultado apresentado.

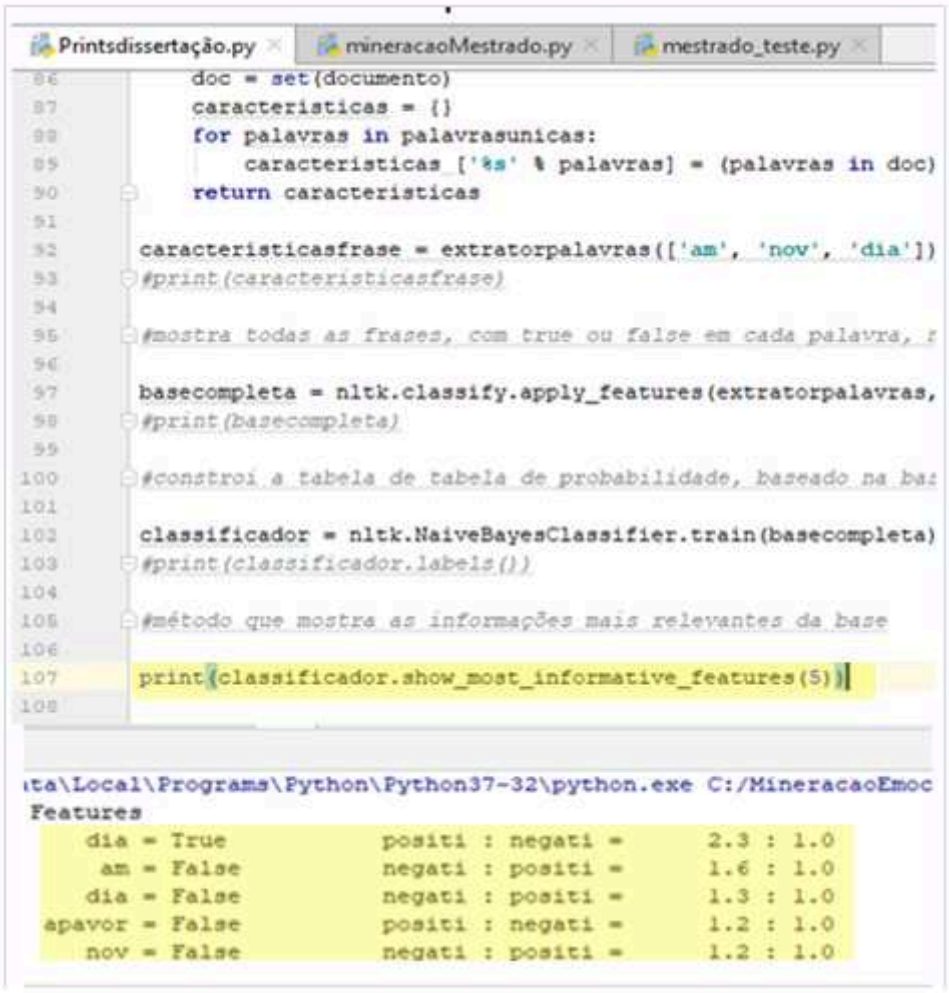
Figura 17 – Código com o método de treinamento e as classes das frases



Fonte: Elaborada pelo autor após a utilização do método labels

Criado também um método classificador.show_mostinformative_features(5) que mostra as palavras mais informativas dentro da base em uma tabela de probabilidade. Nesse estágio, pode ser escolhido o número de palavras que você deseja verificar, nesse caso cinco palavras.

Figura 18 – Aplicação do método mostinformative_features



Fonte: Elaborada pelo autor

Na Figura 18, observam-se as probabilidades, na parte de baixo (assinalado em amarelo). Quando aparece *true* significa que a possibilidade de a frase ser positiva é 2.3 vezes maior do que ser negativa. Quando tem-se *false* a probabilidade de ser uma frase negativa é 1.6 vezes maior do que ser positiva.

Nessa etapa também são repetidas algumas funções ou métodos, para que se possa observar com maior atenção ao funcionamento do código.

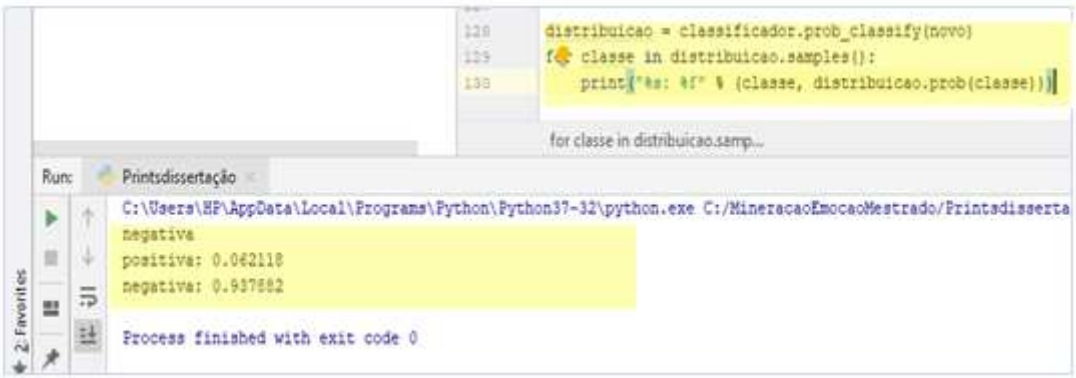
Assim, foi inserida a frase: estou com medo, da base aleatória e realizado todo o pré-processamento novamente. Esse texto também foi inserido na lista de palavras da base. Na tela a seguir, pode-se observar o código, a impressão dos radicais do texto e os textos inseridos na base.

Figura 19 – Inserção de frase na base

Fonte: Elaborada pelo autor

O mesmo texto foi classificado automaticamente pelo algoritmo *naive bayes*, por meio do método *classify*. Foi mostrada também a probabilidade da frase ser positiva ou negativa. Isso pelo método de distribuição de probabilidade.

Figura 20 – Classificação da nova frase



Fonte: Elaborada pelo autor

Pode-se observar na Figura 20 que o algoritmo já consegue classificar como positiva ou negativa uma frase, neste caso classificou como negativa, já consegue também mostrar a porcentagem de positividade (6,3%) ou negatividade (93,7%) de uma frase.

Em seguida, foi implementado o algoritmo, na base de dados extraída da página Novembro Azul, no Facebook®, cujas 645 mensagens foram publicadas de 1º a 30 de

novembro de 2018, Essas mensagens passaram por filtragens, que também fazem parte da etapa de pré-processamento, como, exclusão de frases repetidas e de linhas em branco, correção de palavras abreviadas, exclusão de acentos, de *links*, de caracteres especiais, de frases que não tinham conteúdo relacionado à campanha. Tudo isso antes de usá-las no algoritmo. Depois de todo esse processo de "limpeza", sobraram 261 mensagens, como já relatado anteriormente.

No primeiro teste com a base de dados Novembro Azul, os resultados não foram muito satisfatórios. Pois frases visivelmente negativas, eventualmente são classificadas como positiva pelo algoritmo *naive bayes*. Por exemplo: a frase "Não gostei da campanha" deveria ser classificada como negativa, mas o classificador a marcou como positiva. Podemos observar na Figura 21:

Figura 21 – Frase classificada automaticamente

Fonte: Elaborada pelo autor

5.5 VISUALIZAÇÃO DOS DADOS E AVALIAÇÃO DA APRENDIZAGEM DE MÁQUINA – ETAPA 5

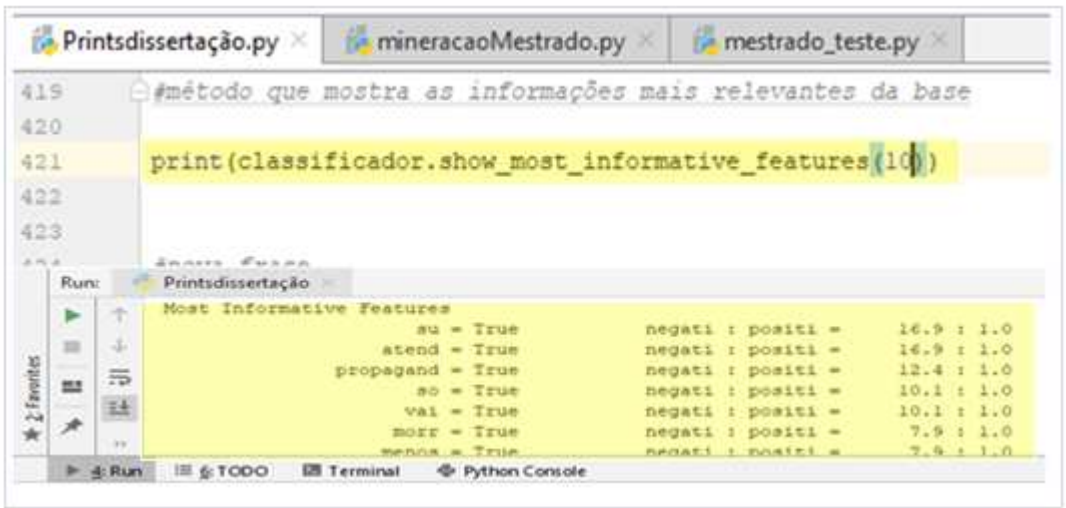
A etapa de análise do modelo criado teve como objetivo testar a acurácia do modelo criado pelo algoritmo *naive bayes*.

Dentro da biblioteca NLTK, existem métodos que realizam essa avaliação. A qual é executada da seguinte maneira: os resultados do modelo de treinamento são comparados com os de teste, criando desse modo uma tabela de erros e de acertos, logo depois, calculando a porcentagem de erros e de acertos do modelo.

A partir dos resultados da acurácia, devemos fazer a avaliação do estudo do cenário do trabalho. Nesse caso, voltado para fins acadêmicos, mas dependendo do contexto onde o modelo será inserido a acurácia deverá ser entre 85% e 100% (MINERAÇÃO..., 2019). O número de classes envolvidas no estudo também influenciam no resultado, quanto mais classes, mais complexa a tarefa de alcançar uma porcentagem alta. No presente estudo, são somente duas classes, mensagens positivas e negativas, então para ser considerada a acurácia, os resultados devem ser altos, ou seja, acima de 85%.

Para auxiliar na visualização de erros ou acertos, foi chamado o método Classificador.show_most_informative_features que mostra as palavras mais significativas da base. Podemos observar na Figura 22 a aplicação do método.

Figura 22 – Palavras mais significativas



Fonte: Elaborada pelo autor

Se fosse detectado algum tipo de erro depois da utilização do método que mostrava as palavras mais significativas, uma possível solução seria adicionar algumas *stopwords* que não

estão na lista da biblioteca NLTK, com o método append. Com o objetivo de testar algumas funcionalidades da biblioteca NLTK, foram adicionadas aleatoriamente duas palavras na lista de *stopwords*, somente para testar o método (vai e tão).

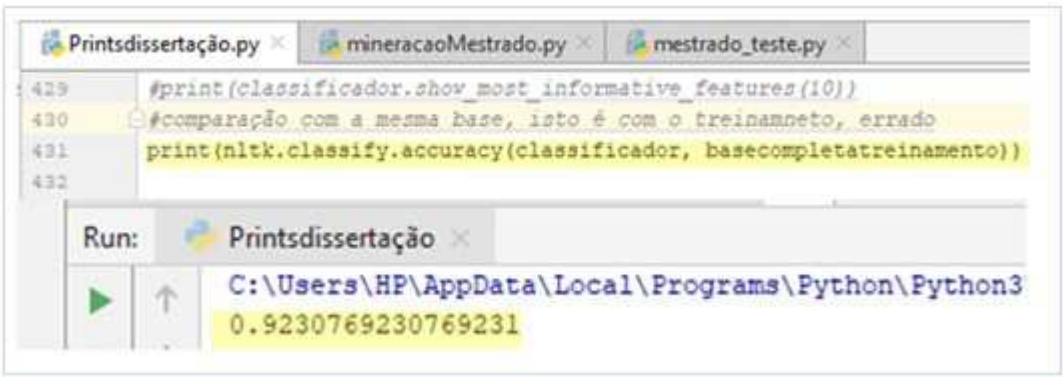
Figura 233 – Inserção de *stopwords*

```
343  #retirando as palavras sem relevância, isto é, que não auxilam no enten
344  stopwords = nltk.corpus.stopwords.words('portuguese')
345  stopwords.append('vai')
346  stopwords.append('tão')
347
348  def removestopwords(texto):
349      frases = []
350      for (palavras, sentimento) in texto:
351          semstop = [p for p in palavras.split() if p not in stopwords]
352          frases.append((semstop, sentimento))
353      return  frases
```

Fonte: Elaborada pelo autor

Procurando melhorar os resultados do algoritmo e testar a sua acurácia, foram feitas algumas implementações no código, principalmente para que se trabalhe com a base de treinamento e com a base de teste. Para isso foi utilizado o método nltk. classify.accuracy, que apresentou ótimo resultado, 92% de acerto. Esse método foi aplicado na mesma base de dados em que ele treinou (base treinamento), somente para testar o método de classificação.

Figura 24 – Teste da acurácia na base treinamento



Fonte: Elaborada pelo autor

Logo a seguir foi alterado o código-fonte e utilizada a base completa de teste, a base recomendada para esse tipo de teste do algoritmo. Chegou-se a um resultado de 84,8%, que

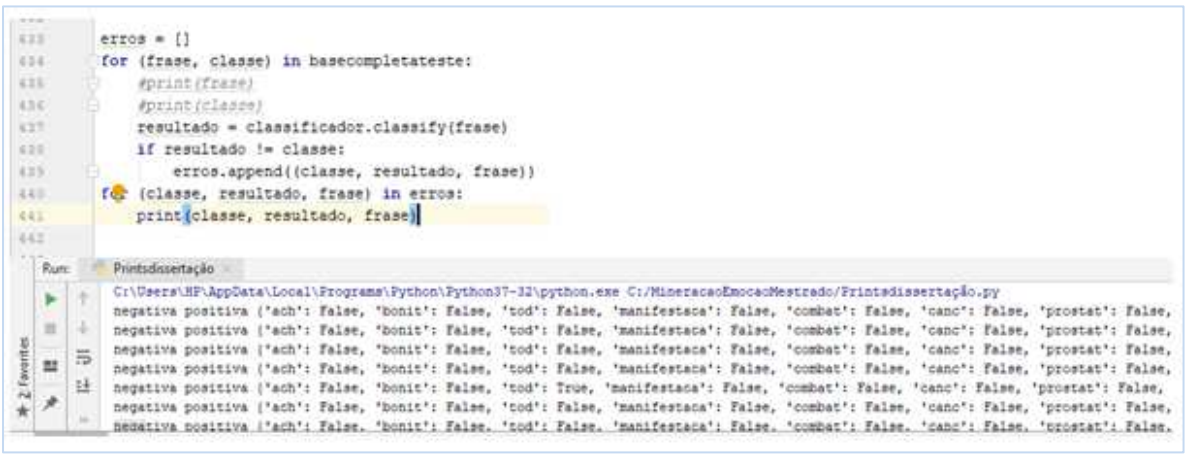
não é considerado ruim, mas existem alguns trabalhos nessa área que têm resultados superiores. Dentro do cenário de frases, comentários ou postagens, tem-se grande dificuldade em classificar, então o resultado é considerado bom, pois de cada 100 mensagens, 84 estariam classificadas corretamente. Pelo número de classes, que são somente duas (positiva a negativa), também é considerado bom o resultado.

Figura 25 – Resultado da acurácia utilizando a base de teste

Fonte: Elaborada pelo autor

Existem maneiras de visualizar de melhor forma o que o classificador apontou como errado ou certo, dessa maneira, medidas podem ser adotadas pelo pesquisador, podendo assim melhorar os resultados. Pode-se observar na Figura 26 o código que mostra a classificação das frases.

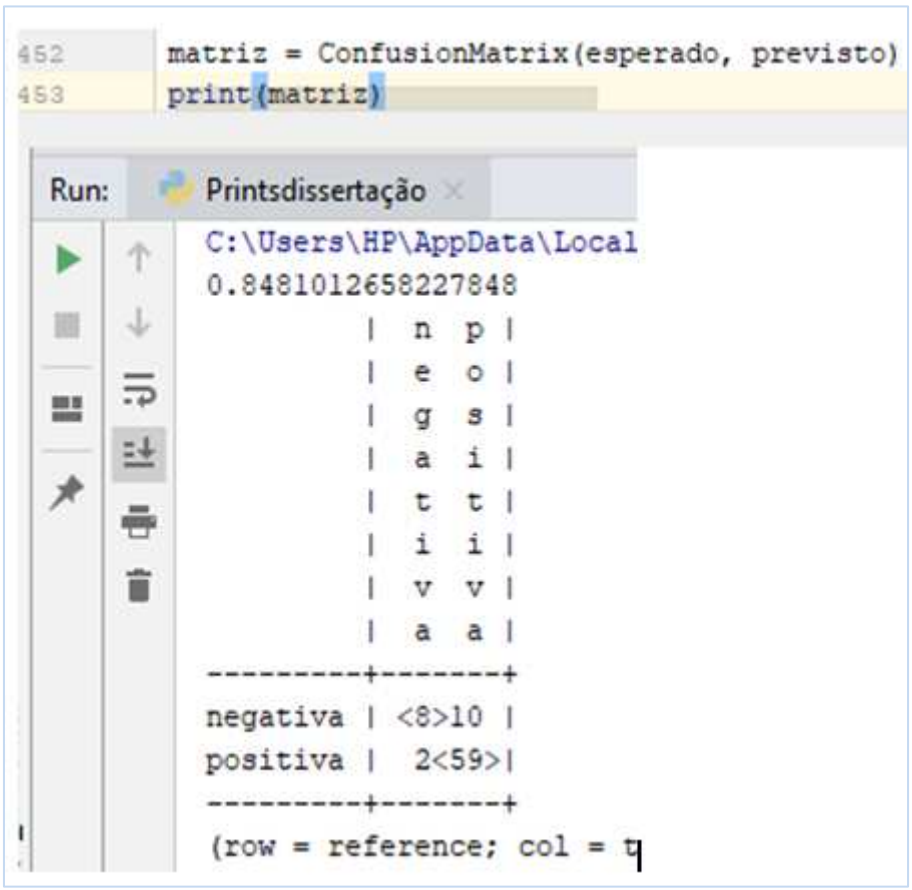
Figura 26 – Classificação da base de teste



Fonte: Elaborada pelo autor

Foi utilizado o método chamado ConfusionMatrix,, que pode auxiliar na forma de visualizar os erros e acertos classificados pelo algoritmo na base de dados. A seguir na Figura 27, tem-se o código e a impressão da matriz de confusão.

Figura 27 – Matriz de confusão



Fonte: Elaborada pelo autor

Na matriz é mostrado dentro dos símbolos de menor e maior (< >) o número de acertos da classificação automática do algoritmo, assim pode-se observar que os maiores erros da classificação automática do algoritmo são na classe negativa, pois num total de 18 frases negativas o algoritmo errou 10, já na classe positiva, acertou 59 frases e errou apenas 2.

6 ANÁLISE DOS RESULTADOS

Nesta seção, apresentam-se os resultados da metodologia aplicada e do desenvolvimento do algoritmo classificador de mensagens. Para tanto se discorre acerca dos seus pontos fortes e fracos.

6.1 ANÁLISE DA ETAPA 1

O processo de extração de dados e criação do *dataset* utilizado neste trabalho é o mais recorrente na literatura, contudo, pode-se destacar que para este estudo foi utilizado para sua elaboração também uma plataforma de apoio, a Netlytic, que realizou a extração dos dados do Facebook®. Já o trabalho nos textos consultados há somente a descrição do uso da API da rede social junto com códigos da linguagem Python, sem o uso de plataforma para intermediar o processo. Optou-se utilizar essa plataforma pelo seu grau de confiabilidade, também por ser uma plataforma gratuita para estudantes e pesquisadores e principalmente por disponibilizar todos os recursos necessários para a coleta de forma automática, como já mencionado.

A plataforma Netlytic, mostrou-se bastante eficiente em relação à disponibilidade da base e aos prazos de entrega. Pois, o serviço foi solicitado/programado no dia 14 de outubro de 2018, para a coleta de 1º a 30 de novembro, os resultados ficaram disponíveis no dia 2 de dezembro no formato CSV, totalizando 645 comentários prontos para serem exportados para trabalhar com um programa de planilha eletrônica. Contudo, foi necessário preceder a limpeza e organização da planilha, retirando as colunas id, guid, link, pubdate, author, author_id, frases neutras sobre a campanha, linhas em branco e linhas repetidas; essa ação foi realizada manualmente pelos pesquisadores, resultando depois da limpeza em 261 comentários. Mas é importante ressaltar que em bases de dados com um número de resultados mais significativos, por exemplo, acima de 1.000 mensagens, é indicado realizar essa atividade automaticamente a partir de comandos da própria planilha eletrônica ou por meio de comandos da linguagem Python, podendo se fazer necessário conhecimentos do programa ou de terceirização do serviço.

No que diz respeito aos critérios do API do Facebook®, entende-se que a restrição da API do Facebook® em relação a não incluir respostas a comentários na base, pode interferir nos resultados, pois “atrapalha” a busca. De qualquer forma, a restrição do API, assim como as demais, está fora do alcance de ações corretivas do pesquisador, pois é uma restrição do

Facebook® com vistas a limitar a extração dos dados e assim proteger a privacidade dos usuários.

6.2 ANÁLISE DA ETAPA 2

A linguagem de programação Python é gratuita, ou seja, está à disposição de todos os pesquisadores, e é uma linguagem livre e multiplataforma. Isso significa que os programas escritos em uma plataforma serão executados sem nenhum problema na maioria das demais existentes sem nenhuma modificação, isto é, computadores com qualquer sistema operacional podem usar a linguagem e, caso a plataforma objetivo não tenha uma versão de Python, desenvolvedores têm a liberdade de estudar e modificar o código da linguagem para fazer com que ela rode onde quer que seja (PYSCIENCE BRASIL, 2019). Considera-se assim um excelente recurso para fins de pesquisa de mineração de textos.

Sobre as *stopwords*, apresentadas no Quadro 1 constituído de todas as *stopwords* do idioma português, da biblioteca NTLK, pertencente à linguagem Python, pode-se observar que além de palavras gramaticais, como preposições e artigos, a biblioteca é composta de advérbios, pronomes possessivos e de verbos auxiliares. Sendo que desses não foi possível identificar as flexões 'houverei'; 'houverão' e 'tém', nem em português brasileiro (PB) ou em português europeu (PE). Assim, seria importante analisar mais aprofundadamente os itens da biblioteca de modo a refiná-la.

Além disso, observou-se que o advérbio 'não', que está presente na biblioteca interferiu em muitos casos de forma negativa na classificação realizada pelo algoritmo. Em algumas frases ele acabou mudando totalmente o sentido da frase. Logo, pode haver a necessidade de que algumas das palavras presentes na NTLK tenham que ser revistas.

Outro filtro utilizado na etapa de pré-processamento é a aplicação do método da biblioteca NLTK denominado de *stemming.* Que deixou somente os radicais das palavras. Contudo, refletiu-se sobre esse comando indagando: se são os prefixos, além de outras funções, são determinantes de negativas, por exemplo: **in**feliz, **in**satisfeito, **in**capaz, **i**napto, **des**cuidado, **im**possível. Talvez fosse importante estabelecer criteriosamente que tipo de advérbios (de ação contrária: **anti**, de afastamento: **ab** etc.) poderiam estar na biblioteca e quais não. Isso vai ao encontro das preocupações apresentadas em Hull e Grenfenstette (1996; FULLER; ZOBEL, 1988 *apud* ALVAREZ, 2014, p. 19) nas quais os pesquisadores manifestavam não ser uma boa ideia usar a técnica de *stemming*.

Já a função que mostra somente as palavras, sem o número de vezes que elas aparecem é importante, pois auxilia o pesquisador a conhecer mais detalhadamente a sua base de dados, suas palavras e suas características. Com o mesmo intuito, aplica-se a função apresentada na Figura 15, processo que vai apontando ao pesquisador se as etapas estão sendo corretamente implementadas, o que é de suma importância.

6.3 ANÁLISE DA ETAPA 3

A classificação humana foi realizada conjuntamente e demandou aproximadamente 9 horas de trabalho. Para uma base consideravelmente maior, por exemplo, 10 vezes o tamanho da Novembro Azul, seria interessante aumentar o número de participantes na equipe e dividir em sub equipes de 3 em 3, dividindo também a quantidade de frases para cada equipe, por exemplo.

A classificação humana foi realizada por profissionais de tecnologia e parece ter resultado satisfatório, contudo, por ter sido realizada sincronicamente não possibilitou observar qual o grau de discordância ou de concordância plena. Para trabalhos futuros talvez fosse importante um tempo maior para discussões sobre a escolha pela classe A ou classe B. Outro fator, que poderia melhorar a pesquisa seria atuar com equipe multidisciplinar, com a participação de pelo menos um linguista, talvez, especificamente, um semanticista, monitorando a classificação com objetivo de verificar se os profissionais envolvidos estão classificando corretamente as frases, demonstrando assim se estão aptos a realizar o seu trabalho de forma eficiente.

6.4 ANÁLISE DA ETAPA 4

A utilização e incremento do algoritmo *naive bayes* foram ações de suma importância ao trabalho, pois esse algoritmo e seus métodos tornam o trabalho de classificação automática mais “fácil” e com bons resultados, todavia, as etapas anteriores devem se realizadas da maneira indicada e com muita atenção aos detalhes.

Esse algoritmo treinou com a base chamada de treinamento, que corresponde a 70% dos dados da base Novembro Azul, implementado com o método chamado NaiveBayes.train, que aprende com o dados já classificados pelos especialistas e classifica os dados automaticamente.

A primeira classificação foi realizada na base de treinamento, o que não é a indicado para a aplicação desse tipo de método, pois, como o algoritmo treinou com essa base, ele teve mais "facilidade" de classificar de maneira correta os dados dessa mesma base. Mas o objetivo nesta etapa foi criar um método que classifique uma base que ainda não foi classificada, então foi modificado o código-fonte e aplicado na base de teste, ou seja, nos 30% restantes da base Novembro Azul.

Outro fator importante desta etapa foi o uso do método classificador.prob_classify, ele mostrou, como se pode observar na Figura 17 a porcentagem de uma frase ser positiva ou negativa.

6.5 ANÁLISE DA ETAPA 5

O teste de acurácia demonstrado na Figura 24 traz um resultado bastante satisfatório, mas faz emergir uma preocupação. Trata-se do fato de que foi utilizada para teste a mesma base de dados do treinamento, a partir de orientações de curso específico[5] para o uso dessas técnicas.

Contudo, durante o estudo, ficou demonstrado que aplicar esse procedimento é um equívoco, e que pode acontecer com frequência; o correto é usar a base de treinamento como referência, mas os testes têm que ser realizados na base de testes. O ideal é sempre utilizar a parte da base correta, nesse caso os 30% de teste (da base do Novembro Azul).

O primeiro teste da acurácia demonstrado na Figura 24 foi muito satisfatório, pois chegou a mais de 92%. Mas como não é recomendado treinar e testar na mesma base de dados, um novo teste foi realizado, aplicando o método na base de testes, conseguindo também um bom resultado, demonstrado na Figura 25, chegando a aproximadamente 85% de acurácia, que é considerado por especialistas da área uma ótima porcentagem.

Tabela 2 – Acurácia das bases

Acurácia base treinamento	**Acurácia base teste**
92,30%	84,81%

Fonte: Elaborada pelo autor

[5] Curso Mineração de Emoção em Textos com Python e NLTK.

Mesmo considerando que os 84,81% já é um bom resultado de classificação automática, foi interessante ter uma melhor visualização dos erros e acertos do método por meio da matriz de confusão mostrado na Figura 28, que demostrou de forma clara que os acertos das classificações automáticas de frases positivas foram bem superiores aos acertos das classificações automáticas de frases negativas, sendo essa ação recomendada para trabalhos futuros.

Em relação à acurácia, no ano de 2015, o modelo criado por Shahana e Omman atingiu uma porcentagem de 83% no teste das frases positivas e 53% nas frases negativas. Números parecidos com o deste trabalho, que demonstram um possível problema na classificação das frases negativas. Já no estudo de Bhaskar *et al*. (2015), os pesquisadores chegaram a um resultado de acurácia de 76% na classificação geral (frases positivas e negativas), ou seja, abaixo do resultado do presente estudo (Tabela 2). Igualmente ocorreu em Kansal e Toshniwal (2014) cujo modelo atingiu aproximadamente 80% de acurácia no geral.

Por meio da utilização da matriz de confusão, teve-se melhor detalhamento do que o algoritmo classificou da forma correta ou incorreta. Das 61 frases classificadas pelos especialistas como positivas, o método acertou 59, isto é, 96,72%, já nas frases negativas a porcentagem de acertos caiu muito, das 18 frases classificadas pelos especialistas como negativas, o algoritmo acertou somente 8, isto é, 44,44%, número muito abaixo do esperado. Isso demostra que método tem uma alta capacidade de classificação correta das frases positivas e por outro lado uma grande dificuldade em classificar frases negativas de forma correta.

Como já destacado anteriormente, um dos fatores que podem influenciar positivamente, nesse caso, seria o melhor tratamento nos prefixos das palavras, ou ainda um trabalho minucioso em algumas palavras consideradas *stopwords*, por exemplo, remoção da palavra “não” da biblioteca.

7 CONSIDERAÇÕES FINAIS

Neste capítulo apresentam-se as contribuições do estudo, suas limitações e recomendações para possíveis trabalhos futuros.

A presente dissertação apresentou a criação e utilização de um modelo classificador automático de polaridade de mensagens, nesse caso da campanha de saúde sobre o câncer de próstata, dos comentários extraídos da página Novembro Azul, do Facebook®.

Os principais resultados da pesquisa demonstram que a utilização do modelo de classificação automática do estudo pode contribuir como uma ferramenta de apoio nos processos de tomada de decisão, principalmente para a equipe gestora ou até mesmo para a equipe de *marketing* do órgão que comanda esse tipo de campanha.

Evidencia-se nos resultados que a etapa de pré-processamento dos textos é muito importante para a classificação automática realizada pelo algoritmo *naive bayes*, mesmo conseguindo quase 85% de acurácia, que é considerado um bom resultado, um melhor foco na etapa de pré-processamento, principalmente no método *stemming* e um melhor trabalho com as *stopwords* podem aumentar consideravelmente os resultados.

Ainda sobre o teste de acurácia, como já destacado no capítulo anterior, o resultado do teste do presente estudo comparado a trabalhos semelhantes, conseguiu resultados muito satisfatórios, obteve resultados superiores, por exemplo, comparado ao trabalho de Shahana e Omman, em 2015, 83%, menor do que o resultado de quase 85% do método criado, já no trabalho de Bhaskar *et al*. (2015), no qual chegaram a 76%, também menor que o método desenvolvido, por fim Kansal e Toshniwal (2014) alcançaram 80%, outro valor abaixo do modelo proposto.

Assim, considera-se que os objetivos específicos foram alcançados, pois:

1. foram selecionadas na literatura e estudadas diversas ferramentas e/ou modelos de extração, análise e classificação automática de textos não estruturados em redes sociais. Ferramentas/modelos que em sua maioria acabaram contribuindo para o trabalho atual, pois trazem muitas técnicas utilizadas no mercado. No capítulo 2, têm-se vários exemplos e aplicações das ferramentas e/ou métodos utilizadas para esse tipo de trabalho
2. Foi desenvolvido e aplicado o modelo de extração, pré-processamento, treinamento, classificação e visualização automatizada nos textos não estruturados da campanha de prevenção ao câncer de próstata. Os capítulos 5 e 6 mostram detalhadamente a criação, aplicação e o uso de técnicas para que

modelo proposto apresente melhores resultados. Exibindo as possíveis falhas e as potencialidades do modelo.

3. Foi mensurada a acurácia do modelo classificador automático de manifestações das emoções. No capítulo 6, têm-se os testes de acurácia do modelo, que aponta boa porcentagem de acertos. Isso indica que o modelo criado poderá ser utilizado como ferramenta auxiliadora nos processos de tomada de decisão dos gestores e profissionais de saúde de campanhas de prevenção de doenças. Têm-se também algumas possíveis sugestões de ajustes, para a melhora do modelo proposto, apresentadas no capítulo sobre resultados.

Considera-se alcançado também o objetivo geral deste estudo, que foi o de *desenvolver e analisar um modelo classificador automático de polaridade de manifestações das opiniões manifestadas por meio de opinião dos seguidores na campanha realizada, em novembro de 2018, sobre o câncer de próstata, pela página denominada Novembro Azul no Facebook®*. Contudo, fez-se necessário transitar os objetivos específicos a favor de criar uma sustentação para propor um modelo forte, baseado na literatura, que possa ser aplicado em diferentes campanhas de saúde. A construção do modelo é demostrada no capítulo 5, seus testes e possíveis ajustes são destacados no capítulo 6.

O trabalho contribui cientificamente cooperando no avanço das pesquisas de ferramentas ou métodos, principalmente os focados no processo de descoberta de conhecimento em textos não estruturados.

O estudo mostrou possíveis limitações, a primeira a ser destacada é que o modelo criado foi moldado para uma campanha de saúde, que possivelmente pode ser utilizado sem muitas alterações em outras campanhas da mesma natureza, mas se for optado por utilizá-lo em outro tipo de área/segmento faz-se necessário algumas adaptações, em relação aos *stemming* e as *spotwords*, pois cada particularidade acaba impactando no desenvolvimento do modelo.

Outra particularidade que pode ter limitado o estudo foi trabalhar sem editar de forma minuciosa a lista de *stopwords* da biblioteca NLTK, recordando que as *stopwords* são as palavras que teoricamente não influenciam no sentido da frase, mas na prática essa premissa precisa ser reavaliada, pois, neste estudo, foram detectadas algumas palavras que estavam na lista que acabaram influenciando de forma negativa nos resultados da classificação automática do modelo criado, no caso a *stopword* ‘não’.

Outro aspecto que pode ter influenciado de forma negativa o estudo foi a utilização do método denominado *stemming* sem tratamento de exceções. É bom frisar que esse método tem a função de deixar somente os radicais das palavras, removendo automaticamente prefixos e sufixos, em algumas situações ele pode ter removido prefixos que podem mudar o sentido das palavras, e também influenciar de forma negativa na classificação automática do modelo proposto.

Outra possível limitação do estudo foi o número de mensagens que foram classificadas, seria interessante buscar uma base de dados com mais mensagens, podendo assim verificar outros aspectos do modelo, por exemplo, o seu desempenho com número grande de dados.

Como trabalho futuro, vislumbra-se a evolução do método, focando principalmente em três adaptações, uma delas na chamada lista de *stopwords*, a outra na etapa de *stemming* e por fim buscar trabalhar com uma base de dados com mais informações. Sugere-se primeiramente um melhor tratamento da lista de *stopwords*, isto é, observar minuciosamente cada uma das palavras da lista, excluir ou adicionar novas palavras conforme o escopo da campanha. Outra evolução importante no método seria um trabalho personalizado na etapa de *stemming*.

Nas três adaptações sugeridas é importante ressaltar que é de suma importância trabalhar com o apoio ostensivo de equipes multidisciplinares, além dos profissionais da área de tecnologia, contar com a colaboração de profissionais da área de psicologia, da área de saúde, da área de linguística, se possível um semanticista.

Por fim, depois de resolver as situações propostas, seria interessante criar uma plataforma *online* ou uma aplicação mobile, com *design* amigável, que realize todo o processo de maneira automática.

Conclui-se que apesar de algumas limitações e indicações de possíveis trabalhos futuros o modelo desenvolvido neste estudo poderá ser utilizado como uma ferramenta auxiliadora nos processos de tomada de decisão dos gestores e profissionais de saúde de campanhas de prevenção de doenças.

A campanha realizada, pela página denominada Novembro Azul no Facebook®, em novembro de 2018, parece poder ser considerada como positiva, uma vez que 77% das mensagens validas foram classificadas pelos especialistas em tecnologia como positivas.

REFERÊNCIAS

AGNIHOTRI, R. *et al*. Social media: influencing customer satisfaction in B2B sales. **Industrial Marketing Management**, EUA, v. 53, p. 172–180, 2016.

ALVARES, R. V. **Algoritmos de Stemming e o estudo de proteomas**. 2014. 82 f. Tese (Doutorado em Engenharia de Sistemas e Computação) – Programa de Pós-Graduação em Engenharia de Sistemas e Computação (COPPE). Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

ARAUJO, G. D. **Análise de sentimento de mensagens do twitter em português brasileiro relacionadas a temas de saúde**. 2014. 84 f. Dissertação (Mestrado em Ciências) – Programa de Pós-Graduação em Gestão e Informática em Saúde. Universidade Federal de São Paulo, Escola Paulista de Medicina, São Paulo, 2014.

BAE, Y.; LEE, H. Sentiment analysis of twitter audiences: measuring the positive or negative influence of popular twitterers. **Journal of the American Society for Information Science and Technology**, New York, v. 63, n. 12, p. 2.521–2.535, 2012.

BERRY, M. W.; KOGAN, J. **Text mining**: applications and theory. [*S.l.*]: John Wiley & Sons, 2010.

BHASKAR, Jasmine; SRUTHI, K.; NEDUNGADI, Prema. Hybrid approach for emotion classification of audio conversation based on text and speech mining. **Procedia Computer Science**, v. 46, p. 635-643, 2015.

BRASIL. **Lei n. 13.709, de 14 de agosto de 2018**. Dispõe sobre a proteção de dados pessoais e altera a Lei n. 12.965, de 23 de abril de 2014 (Marco Civil da Internet). Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm. Acesso em: 7 jul. 2019.

BUTLER ANALYTICS. **Text Analytics**: a business guide. 2014. Disponível em: http://www.butleranalytics.com/wp-content/uploads/2014/02/Text-AnalyticsGuide.pdf. Acesso em: 17 maio 2019.

CÁCERES, J. G. Las nuevas tecnologías de información y comunicación y las políticas culturales en México: ingeniería en comunicación social del servicio de redes sociales Facebook. **Intercom**, São Paulo, v. 34, p. 175–196, 2011.

CAIRE, L. F. Hipnose em pacientes oncológicos: um estudo psicossomático em pacientes com câncer de próstata. **Psico-USF**, [*s.l.*], v. 17, n. 1, p. 153–162, 2012.

CAMBRIA, E. Affective computing and sentiment analysis. **IEEE Intelligent Systems**, EUA, v. 31, n. 2, p. 102–107, 2016.

CARLEY, K. M. *et al*. Crowd sourcing disaster management: the complex nature of Twitter usage in Padang Indonesia. **Safety science**, EUA, v. 90, p. 48–61, 2016.

CERQUEIRA, A. Jr. *et al*. **Implementação de buscas utilizando linguagem natural através de algoritmos adaptativos**. 2010. Disponível em: https://slideplayer.com.br/slide/3679696/. Acesso em: 4 jun. 2019.

CHAN, E. H. *et al*. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. **PLoS neglected tropical diseases**, EUA, v. 5, n. 5, e1206, 2011.

CHENG, Y. *et al*. Mineração de dados e conhecimento com grandes volumes de dados para produção inteligente. **Revista de Integração da Informação Industrial**, [*s.l.*], v. 9, p. 1–13, 2018.

CHEONG, M.; LEE, V. C. S. A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter Inf. **Syst. Front**., v. 13, n. 1, p. 45–59, 2011.

CHEW C.; EYSENBACH, G. Pandemics in the Age of Twitter: content analysis of tweets during the 2009 H1N1 Outbreak. **PLoS ONE**, Toronto, Canada, v. 5, n. 11, e14118, 2010.

CONCEITOS. **Rede Social**. [2019]. Disponível em: https://conceitos.com/rede-social. Acesso em: 24 maio 2019.

CRANNELL, W.C. *et al*. A pattern-matched twitter analysis of US cancer-patient sentiments. **J. Surg. Res.**, [*s.l.*], v. 206, n. 2, p. 536–542, 2016.

DA SILVA. Rafael Rodrigues. **Brasil é o segundo país do mundo a passar mais tempo na internet**. 2019. Disponível em: https://canaltech.com.br/internet/brasil-e-o-segundo-pais-do-mundo-a-passar-mais-tempo-na-internet-131925/. Acesso em: 2 out. 2019.

DAI, H.; HAO, J. Mining social media data for opinion polarities about electronic cigarettes. **Tobacco control**, [*s.l.*], v. 26, n. 2, p. 175–180, 2017.

DE ARAÚJO, G. D. *et al*. Análise de sentimentos sobre temas de saúde em mídia social. **Journal of Health Informatics**, São Paulo, v. 4, n. 3, p. 1–5, jul./set. 2012.

DE BRITO, E. M. N. **Mineração de Textos**: detecção automática de sentimentos em comentários nas mídias sociais. 2017. 88 f. Dissertação (Mestrado em Sistemas de Informação e Gestão do Conhecimento) – Programa de Pós-Graduação em Stricto Sensu, Sistema de Informação e Gestão do Conhecimento da Universidade Fundação Mineira de Educação e Cultura. Belo Horizonte, 2017.

DE FRANÇA, T. C.; OLIVEIRA, J. Análise de sentimento de tweets relacionados aos protestos que ocorreram no Brasil entre junho e agosto de 2013. *In*: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAN), 3., 2014, [*s.l.*]. **Proceedings** [...], [*s.l.*:*s.n*.], 2014. p. 128–139.

EXAMES. **Facebook chega a 125 milhões de usuários no Brasil**. 2018. Disponível em: https://exame.abril.com.br/tecnologia/Facebook-chega-a-125-milhoes-de-usuarios-no-brasil/. Acesso em: 27 maio 2019.

FARZINDAR, A.; INKPEN, D. Natural language processing for social media. **Synthesis Lectures on Human Language Technologies**, [*s.l.*], v. 8, n. 2, p. 1-166, 2015.

FELDMAN, R. *et al*. Knowledge Management: a text mining approach. *In*: INT. CONF. ON PRACTICAL ASPECTS OFKNOWLEDGE MANAGEMENT (PAKM98), 2., Basel. Switzerland, 29 a 30 oct. 2017. **Proccedings** […], Basel. Switzerland, 2017. p. 1–10. Disponível em: https://pdfs.semanticscholar.org/47cc/1519218acdd7693683fbf47a6c103a63ccfa.pdf. Acesso em: 30 ago. 2019.

FORTUNY, E. J. D. *et al*. Media coverage in times of political crisis: a text mining approach. **Expert Systems with Applications**, New York, v. 39, n. 14, p. 11. 616–11.622, 2012.

GABARDO, A. C. **Análise de rede social**: uma visão computacional. São Paulo: Novatec Editora, 2015.

GIATSOGLOU, M. *et al*. Sentiment analysis leveraging emotions and word embeddings. **Expert Systems with Applications**, New York, v. 69, p. 214–224, 2017.

GHORPADE, A. S. Intelligent data mining of social media for improving health care. International. **Journal of Innovative Research in Science, Engineering and Technology**, [*s.l.*], v. 6, issue 8, Aug. 2017. Disponível em: http://www.ijirset.com/upload/2017/august/64_paper2%20_7_.pdf. Acesso em: abr. 2019.

GOMIDE, J. *et al*. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. *In*: OF THE 3RD INTERNATIONAL WEB SCIENCE CONFERENCE. ACM, 2011, [*s.l.*]. **Proceedings** [...], [*s.l.*]: ACM, 2011. p. 3.

GONÇALVES, P. *et al*. Comparing and combining sentiment analysis methods. *In*: ACM CONFERENCE ON ONLINE SOCIAL NETWORKS, 1., 2013,[*s.l.*]. **Proceedings** […], [s.*l.*]: ACM, 2013. p. 27–38.

GRUZD, A.; PAULIN, D.; HAYTHORNTHWAITE, C. Analyzing social media and learning through content and social network analysis: A faceted methodological approach. **Journal of Learning Analytics**, [*s.l.*], v. 3, n. 3, p. 46–71, 2016.

HU, Y.-H.; CHEN, Y.-L.; CHOU, H.-L. Opinion mining from online hotel reviews–a text summarization approach. **Information Processing & Management**, [*s.l.*], v. 53, n. 2, p. 436–449, 2017.

INJADAT, M. N.; SALO, F.; NASSIF, A. B**. Data Mining Techniques in Social Media**: a survey data mining techniques in social media: a survey, neurocomputing. 2017. Disponível em: https://www.researchgate.net/publication/304401202_Data_Mining_Techniques_in_Social_Media_A_Survey. Acesso em: abr. 2019

INSTITUTO LADO A LADO PELA VIDA. **Portal virtual**. 2019. Disponível em: http://www.ladoaladopelavida.org.br/. Acesso em: 27 maio 2019.

INSTITUTO NACIONAL DE CÂNCER (INCA). **Estimativa 2018**. Incidência de Câncer no Brasil. 2018. Disponível em: http://www1.inca.gov.br/estimativa/2018. Acesso em: 4 jun. 2019.

ISAH, H.; TRUNDLE, P.; NEAGU, D. **Social media analysis for product safety using text mining and sentiment analysis**. Bradford, UK: IEEE, 2014. p. 1–7.

JAYAMALINI, K.; PONNAVAIKKO, M. Research on web data mining concepts, techniques and applications. *In*: INTERNATIONAL CONFERENCE ON ALGORITHMS, METHODOLOGY, MODELS AND APPLICATIONS IN EMERGING TECHNOLOGIES (ICAMMAET). [*S.l.*]: IEEE, 2017. p. 1–5.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing**: an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N. J.: Pearson Prentice Hall, 2009.

KANSAL, H.; TOSHNIWAL, D. Aspect based summarization of context dependent opinion words. **Procedia Computer Science**, [*s.l.*], v. 35, p. 166-175, 2014.

KAVANAUGH, A. L. *et al*. Social media use by government: from the routine to the critical. **Government Information Quarterly**, Amsterdam, v. 29, n. 4, p. 480–491, 2012.

KHARYA, S. Using data mining techniques for diagnosis and prognosis of cancer disease. **International Journal of Computer Science**, Engineering and Information Technology (IJCSEIT), Chhattisgarh, India, v. 2, n. 2, p. 1–12, 2012.

KIETZMANN, J. H. *et al*. Mídia social? Fale sério! Compreender os blocos de construção funcionais das mídias sociais. **Horizontes de Negócios**, Rio de Janeiro, v. 54, n. 3, p. 241–251, 2011.

KONTOPOULOS, E. *et al*. Ontology-based sentiment analysis of twitter posts. **Expert Systems with Applications**, New York, v. 40, n. 10, p. 4.065–4.074, 2013.

KUMAR, V.; MIRCHANDANI, R. Aumentando o ROI do marketing de mídia social. **MIT sloan management review**, [*s.l.*], v. 54, n. 1, p. 55, 2012.

LIU, B. Sentiment analysis: a multi-faceted problem. **The IEEE Intelligent Systems**, [*s.l.*], v. 25, p. 1–5, 2010.

LIU, B.; ZHANG, L. A survey of opinion mining and sentiment analysis. *In*: AGGARWAL, C. C.; ZHAI, C. (ed.). **Mining text data**. Chicago: Springer, 2012. cap. 13, p. 415–463.

LUCCA, G. *et al*. **Uma implementação do algoritmo naıve bayes para classificação de texto**. 2013. Disponível em: http://www.lbd.dcc.ufmg.br/colecoes/erbd/2013/0019.pdf. Acesso em: 4 jun. 2019.

MÄNTYLÄ, M. V.; GRAZIOTIN, D.; KUUTILA, M. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. **Computer Science Review**, [*s.l.*], v. 27, p. 16-32, 2018.

MARAGOUDAKIS, M.; LOUKIS, E.; CHARALABIDIS, Y. A review of opinion mining methods for analyzing citizens' contributions in public policy debate. *In*: INTERNATIONAL CONFERENCE ON ELECTRONIC PARTICIPATION. Springer, Berlin: Heidelberg, 2011. p. 298–313.

MARTELETO, R. M. rede social, mediação e apropriação de informações: situando campos, objetos e conceitos na pesquisa em Ciência da Informação. **Revista Brasileira Ciência da Informação**, Brasília, DF, v. 3, n. 1, p. 27–46, jan./dez. 2010.

MEDEIROS, A. P.; MENEZES, M. F. B.; NAPOLEÃO, A. A. Fatores de risco e medidas de prevenção do câncer de próstata: subsídios para a enfermagem. **Revista Brasileira de Enfermagem**, Brasília, DF, v. 64, n. 2, p. 385–388, 2011.

MEJOVA, Y.; WEBER, I.; MACY, M. W. (ed.). **Twitter**: a digital socioscope. Cambridge: Cambridge University Press, 2015.

MENEZES, L. N. *et al*. Conhecimento dos homens com idade acima de 40 anos sobre o câncer de próstata, frequentadores de um ambulatório de especialidade médica. **Hórus**, São Paulo, v. 8, n. 2, p. 11–20, 2017.

MICHAELIDOU, N.; SIAMAGKA, N. T.; CHRISTODOULIDES, G. Uso, barreiras e medição do marketing de mídia social: uma investigação exploratória de pequenas e médias marcas B2B. **Gestão de Marketing Industrial**, [*s.l.*], v. 40, n. 7, p. 1.153–1.159, 2011.

MINISTÉRIO DA SAÚDE. **Câncer de próstata**: causas, sintomas, tratamentos, diagnóstico e prevenção. [2019]. Disponível em: http://portalms.saude.gov.br/saude-de-a-z/cancer-de-prostata. Acesso em: 25 maio 2019.

MINISTÉRIO DA SAÚDE. **Resolução n. 466, de 12 de dezembro de 2012**. O Plenário do Conselho Nacional de Saúde em sua 240ª Reunião Ordinária, realizada nos dias 11 e 12 de dezembro de 2012, no uso de suas competências regimentais e atribuições conferidas pela Lei n. 8.080, de 19 de setembro de 1990, e pela Lei n. 8.142, de 28 de dezembro de 1990. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/cns/2013/res0466_12_12_2012.html. Acesso em: 3 jun. 2019.

MINERAÇÃO de Emoção em Textos com Python e NLTK. [2019]. [curso *online*]. Disponível em: https://www.udemy.com/course/mineracao-de-emocao-em-textos-com-python-e-nltk/learn/lecture/7317124#overview. Acesso em: 30 ago. 2019.

MINUTO SAUDÁVEL. **O que é o Novembro Azul, como surgiu, objetivo e como participar**. [2018]. Disponível: https://minutosaudavel.com.br/novembro-azul/. Acesso em: 19 nov. 2018.

MODESTO, A. A. Dall'Agnol *et al*. Um novembro não tão azul: debatendo rastreamento de câncer de próstata e saúde do homem. **Interface-Comunicação, Saúde, Educação**, Botucatu, v. 22, n. 64, p. 251–252, 2018.

MONTOYO, A.; MARTÍNEZ-BARCO, P.; BALAHUR, A. Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. **Decision Support Systems**, Amsterdam, v. 53, n. 4, p. 675–679, 2012.

MOSTAFA, M. M. More than words: social networks' text mining for consumer brand sentiments. **Expert Systems with Applications**, New York, v. 40, n. 10, p. 4241–4251, 2013.

NETLYTIC. **About**. [2019]. Disponível: https://netlytic.org/home/?page_id=10834. Acesso em: 30 maio 2019.

NGUYEN, T. T.; KRAVETS, A. G. Analysis of the social network facebook comments. *In*: INTERNATIONAL CONFERENCE ON INFORMATION, INTELLIGENCE, SYSTEMS & APPLICATIONS (IISA), 7th., 2016. IEEE, 2016. p. 1-5.

NIELSEN COMPANY. **State of the media**: the social media report 2012. 2012. Disponível em: https://www.nielsen.com/us/en/insights/news/2012/social-media-report-2012-social-media-comes-of-age.html. Acesso em: 4 jun. 2019.

NOVEMBRO AZUL. Facebook: **@NovembroAzulBrasil**. [2019]. Disponível em: https://www.Facebook.com/NovembroAzulBrasil/. Acesso em: 9 jul. 2019.

OHANA, B.; TIERNEY, B. Sentiment classification of reviews using SentiWordNet. *In*: IT&T CONFERENCE, DUBLIN INSTITUTE OF TECHNOLOGY, 22-23 octuber, 9th,. Dublin, Ireland, 2009. p. 1–9.

O'CONNOR, B. *et al*. From tweets to polls: linking text sentiment to public opinion time series. *In*: FOURTH INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 2010, [*s.l.*]. **Proccedings** […], [*s.l.*:*s.n*]. 2010.

OLIVEIRA, G. H. M.; WELCH, E. W. Social media use in local government: linkage of technology, task, and organizational context. **Government Information Quarterly**, Amsterdam, v. 30, n. 4, p. 397–405, 2013.

PARK, S. J. *et al*. Expanding the presidential debate by tweeting: the 2012 presidential election debate in South Korea. **Telematics and informatics**, [*s.l.*], v. 33, n. 2, p. 557-569, 2016.

REECE, A. G.; DANFORTH, C. M. Instagram photos reveal predictive markers of depression. **EPJ Data Science**, v. 6, n. 1, p. 15, 2017.

PURWARIANTI, A. *et al*. InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification. *In*: INTERNATIONAL CONFERENCE ON ADVANCED INFORMATICS: concepts, theory and application (ICAICTA), 16 a 19 aug. 2016. George Town, Malaysia: IEEE, 2016. p. 1–5.

PYSCIENCE BRASIL. **O que é Python**? [2019]. Disponível em: http://pyscience-brasil.wikidot.com/python:python-oq-e-pq. Acesso em: 15 ago. 2019.

RAORANE, A.; KULKARNI, R. V. **Data mining techniques**: a source for consumer behavior analysis. 2011. Disponível em: https://arxiv.org/abs/1109.1202 Acesso em: 4 jun. 2019.

RIGHTNOW. **Portal virtual**. [2019]. Disponível em: http://www.rightnow.com. Acesso em: 3 jun. 2019.

RODRIGUES, R. G. **SentiHealth-Cancer**: uma ferramenta de análise de sentimento para ajudar a detectar o humor de pacientes de câncer em uma rede social online. 144 f. Dissertação (Mestrado em Computação) – Programa de Pós–Graduação do Instituto de Informática. Universidade Federal de Goiás. Goiás, 2016. Disponível em: https://pdfs.semanticscholar.org/73bf/80f04813f53c20e59f79de4e7254f5b86687.pdf. Acesso em: 29 jul. 2019.

ROSA, J. L. G. **Fundamentos da Inteligência Artificial**. Rio de Janeiro: LTC, 2011.

SABERI, Bilal; SAAD, Saidah. Sentiment analysis or opinion mining: A review. **IJASEIT**, v. 7, n. 5, 2017.

SALLOUM, S. A. *et al*. A survey of text mining in social media: Facebook and twitter perspectives. **Adv. Sci. Technol. Eng. Syst. J**, EUA, v. 2, n. 1, p. 127–133, 2017.

SARKER, A.; GONZALEZ, G. Data, tools and resources for mining social media drug chatter. *In*: WORKSHOP ON BUILDING AND EVALUATING RESOURCES FOR BIOMEDICAL TEXT MINING (BioTxtM2016), 5., 2016, [*s.l.*]. **Proceedings** [...], [*s.l.*:*s.n.*]. 2016. p. 99–107.

SERRANO-GUERRERO, J. *et al*. Sentiment analysis: A review and comparative analysis of web services. **Information Sciences**, [*s.l.*], v. 311, p. 18–38, 2015.

SILVA, Luciana Kraemer da et al. Análise de Sentimento pela ótica da abordagem multimodal. **CINTED-UFRGS Novas Tecnologias na Educação**, Porto Alegre, v. 15, n. 1, 2017.

SILVA, E. **Técnicas de Data e Text Mining para anotação de um arquivo digital**. 2010. 96 f. Dissertação (Mestrado em Engenharia Eletrónica e Telecomunicações) – Programa de Pós-Graduação em Especialização Sistemas de Informação. Universidade de Aveiro, Portugal, 2010.

SHAHANA, P. H.; OMMAN, Bini. Evaluation of features on sentimental analysis. **Procedia Computer Science**, v. 46, p. 1585-1592, 2015.

SOBKOWICZ, P.; KASCHESKY, M.; BOUCHARD, G. Opinion mining in social media: modeling, simulating, and forecasting political opinions in the web. **Government Information Quarterly**, Amsterdam, v. 29, n. 4, p. 470–479, 2012.

STATISTA. **Number of social media users worldwide from 2010 to 2020**. 2017. Disponível em: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. Acesso em: 27 maio 2019.

THOMAS, M.; PANG, B.; LEE, L. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *In*: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING. 2006, [*s.l.*]. **Proceedings [**...], [*s.l.*]: Association for Computational Linguistics, 2006. p. 327–335.

TONIN, S. **Como surgiu o novembro azul e por que ele é tão importante**? out. 2017. Disponível em: https://sobrebarba.com.br/blogs/blog/o-que-e-o-novembro-azul-e-por-que-ele-e-tao-importante. Acesso em: 19 nov. 2018.

VINODHINI, G.; CHANDRASEKARAN, R. M. Sentiment analysis and opinion mining: a survey. International **Journal of Advanced Research in Computer Science and Software Engineering**, [*s.l.*], v. 2, n. 6, p. 283–292, 2012.

WILSON, H. J. *et al*. Qual é a sua estratégia de mídia social? **Harvard Business Review**, Canada, p. 1–4, ago. 2011. Disponível em: https://hbr.org/2011/07/whats-your-social-media-strategy. Acesso em: 4 jun. 2019.

YANG, D. H.; YU, G. A method of feature selection and sentiment similarity for Chinese micro-blogs. **Journal of Information Science**, Cambridge, v. 39, n. 4, p. 429–441, 2013.

YATES, D.; PAQUETTE, S. Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake. **International journal of information management**, [*s.l.*], v. 31, n. 1, p. 6–13, 2011.

YOON, B.; PARK, Y. A text-mining-based patent network: analytical tool for high-technology trend. **The Journal of High Technology Management Research**, [*s.l.*], v. 15, n. 1, p. 37–50, 2004.

YU, Y.; DUAN, W.; CAO, Q. The impact of social and conventional media on firm equity value: a sentiment analysis approach. **Decision Support Systems**, Amsterdam, v. 55, n. 4, p. 919–926, 2013.

WANG, H. *et al*., 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 50., 8 a 14 jul. 2012, Jeju, Republic of Korea. **Proceedings** [...], 2012. p. 115–120. Disponível em: https://www.aclweb.org/anthology/P12-3020. Acesso em: 30 ago. 2019.

ZHANG, W.; XU, H.; WAN, W. Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. **Expert Systems with Applications**, [*s.l.*], v. 39, n. 11, p. 10.283–10.291, 2012.

ANEXO A – BASE ALEATÓRIA

Base aleatória e suas classificações

[('eu sou admirada por muitos','positiva'),
('me sinto completamente amado','positiva'),
('amar e maravilhoso','positiva'),
('estou me sentindo muito animado novamente','positiva'),
('eu estou muito bem hoje','positiva'),
('que belo dia para dirigir um carro novo','positiva'),
('o dia estÃ¡ muito bonito','positiva'),
('estou contente com o resultado do teste que fiz no dia de ontem','positiva'),
('o amor e lindo','positiva'),
('nossa amizade e amor vai durar para sempre', 'positiva'),
('estou amedrontado', 'negativa'),
('ele esta me ameacando a dias', 'negativa'),
('isso me deixa apavorada', 'negativa'),
('este lugar e apavorante', 'negativa'),
('se perdermos outro jogo seremos eliminados e isso me deixa com pavor', 'negativa'),
('tome cuidado com o lobisomem', 'negativa'),
('se eles descobrirem estamos encrencados', 'negativa'),
('estou tremendo de medo', 'negativa'),
('eu tenho muito negativa dele', 'negativa'),
('estou com medo do resultado dos meus testes', 'negativa')]

APÊNDICE A – FRASES CLASSIFICADAS POR ESPECIALISTAS, BASE TREINAMENTO E BASE TESTE

Base de treinamento

('acho muito bonita toda essa manifestacao de combate ao cancer de prostata pena que nao passe de fitinhas azuis nas camisas jalecos e postos de saude nao ha uma acao efetiva por parte do governo de disponibilizar aos homens de baixa renda exames de psa e muito menos hospitais conveniados com o sus para cirurgias lamentavel',
'negativa'),
('eu tenho cancer de prostata estagio a anos no comeco fiz prostatectomia ai meu psa quase zerou anos depois ele recidivou agora ja fiz radioterapias estou aguardando o novo psa espero com a ajuda de deus e do medico maravilhoso e hospital do sus tambem maravilhoso que ele me conseguiu se eu for merecedor consiga controlar a doenca se cuidem facam o exame de toque nao doi nada e salva vidas tem muitos homens morrendo por causa disso la no hospital',
'positiva'),
('ola homens novembro chegou azul se cuidem prevencao sempre e bom', 'positiva'),
('o diagnostico precoce salva vidas vi isso com meu pai e ele desde quando comecou a fazer os exames ele nunca reclamou porque sabia que sao importantes',
'positiva'),
('faco todo ano o exame comecei no psa depois fiz o de toque meus amigos e necessario fazer seja homem faca o exame da prostata para o bem da sua vida',
'positiva'),
('aconselho que faca tenho 60 anos nunca fiz e nem vou fazer e sim farei quando todas mulheres podem fazer o exame de mama que e so propaganda',
'negativa'),
('bom dia e muito importante que os homens facam a prevencao do exame de prostata tem psa pelo sangue e tem exame de toque que procura acha quem acha a cura e bem mais rapida e menos sofrimento prevencao',
'positiva'),
('ve se pode isso arnaldo estou com sete meses esperando fazer um psa pelo sus',
'negativa'),
('importante prevencao e fundamental', 'positiva'),
('com certeza tem que ser feito esse exame', 'positiva'),
('aonde eles fazem quanta hipocrisia', 'positiva'),
('e isso mesmo juiz arnaldo cesar coelho', 'positiva'),
('valeu arnaldo voce e o cara', 'positiva'),
('a melhor forma de prevencao a qualquer cancer e a alimentacao saudavel', 'positiva'),
('pra quem tem a idade de 40 anos ou mais e melhor se cuidar', 'positiva'),
('e muito importante que os homens cuide da sua saude', 'positiva'),
('top', 'positiva'),
('homens se cuidem', 'positiva'),
('gesto muito nobre', 'positiva'),
('verdade o pessoal de renda baixa nao existe para os governos se nao tiver o dinheiro simplesmente nao faz',
'negativa'),

('estamos juntos vou fazer meu exame da prostata dia 14 de novembro de 2018 a aqui em cotia sao paulo parabens a todos que se preocupam essa doenca e maldosa e traicoeira todos tem que prevenir muito obrigado ate mais boa sorte hohohoho',
'positiva'),
('parabens pela atitude todos homens tem que fazer nao so o psa mais o toque tambem ja faco todo ano cuidar de si e amar a voce proprio',
'positiva'),
('preciso fazer o meu tenh anos ja e o tempo para iniciar um exame da prostata', 'positiva'),
('l e g a l', 'positiva'),
('previna-se', 'positiva'),
('vamos nos prevenir abrace essa ideia', 'positiva'),
('ja rodei no rio de janeiro nao estao fazendo esta dificil', 'negativa'),
('tao esperando o que bora la', 'positiva'),
('como fazer se fizer o exame e constatar a doenca e se nao tiver condicoes de pagar como fazer ja que tratamento requer pressa e sus demora muito agilizar',
'negativa'),
('governo gasta um dinheirao com a propaganda e esquece de pagar os medicos e laboratorios coisa para ingles ver',
'negativa'),
('hipocrisia cade a facilidade pra fazer exames quando se tem a sorte de ter um encaminhamento tem que varar a madrugada pra ter o suposto atendimento vergonha',
'negativa'),
('muita propaganda nenhum empenho esse novembro azul e para quem pode pagar uma consulta que esta em torno de reais fora os exames de sangue e utrasografia isso eu faco todos os anos nao preciso de propaganda agora e os pobres se nao tem dinheiro nem pra comer faz o que esses hospitais estao um caos o eu quero dizer que pobre mesmo esses nao tem vez pra mim essa propaganda e so para os que podem porque os governos nao estao nem ai para os pobres',
'negativa'),
('oi que vai acontecer e simples vao morrer muitos por nao ter atendimento e que o tratamento e dificil e caro',
'negativa'),
('muita propaganda na hora do atendimento meses para passar por um especialista brincadeira', 'negativa'),
('eu tambem concordo com voce e do jeito que voce falou ai mesmo na nossa saude ta um porcaria', 'negativa'),
('esse cidadao deve ter um bom plano de saude quem depender da saude publica morre sem atendimento', 'negativa'),
('se o exame for pelo sus o camarada morre e o resultado nao chega', 'negativa'),
('diagnosticar e a melhor parte depois e pedreira', 'positiva'),
('vamos amigos o medico urologista quer ter um dedo de prosa com a gente', 'positiva'),
('acorda coroada alerta bom do arnaldo', 'positiva'),
('amarildo acorda pra vida', 'positiva'),
('cleiton voce ja foi levar sua dedadinha', 'positiva'),
('so conversa sem acao nenhuma', 'negativa'),
('eu fiz', 'positiva'),
('estamos juntos com todos novembro azul', 'positiva'),
('ola homens novembro chegou azul se cuidem prevencao sempre e bom', 'positiva'),
('o diagnostico precoce salva vidas vi isso com meu pai e ele desde quando comecou a fazer os exames ele nunca reclamou porque sabia que sao importantes',

'positiva'),
('agradeco a todo corpo clinico e a todos da nossa ubs principalmente ao nosso secretario da saude joao victor barboza por estar fazendo tratamento da hiperplasia da prostata em um hospital de referencia hospital euriclides de jesus zerbini o qual quando era vivo tive o prazer de conhece-lo pessoalmente desta forma peco a todos que nunca se submeteram a exames e que estao na idade necessaria para tal nao esquecam de procurar a ubs nossa que sempre faz o possivel para nos encaminhar para fazelos',
'positiva'),
('usam mais agrotoxicos veneno na producao de alimentos e ai como forma de remedios fazem essas campanhas inutil',
'negativa'),
('faco todo ano o exame comecei no psa depois fiz o de toque meus amigos e necessario fazer seja homem faca o exame da prostata para o bem da sua vida',
'positiva'),
('aconselho que faca, tenho 60 anos nunca fiz e nem vou fazer e sim farei quando todas mulheres podem fazer o exame de mama que e so propaganda',
'positiva'),
('jogar a saude para escanteio e falta pra cartao cartao azul aproveite o dia mundial de combate ao cancer de prostata e compartilhe essa advertencia com quem voce se importa afinal e sempre bom lembrar a vida nao e um jogo novembroazul somostodosladoalado',
'positiva'),
('ola homens novembro chegou azul se cuidem prevencao sempre e bom', 'positiva'),
('o diagnostico precoce salva vidas vi isso com meu pai e ele desde quando comecou a fazer os exames ele nunca reclamou porque sabia que sao importantes',
'positiva'),
('faco todo ano o exame comecei no psa depois fiz o de toque meus amigos e necessario fazer seja homem faca o exame da prostata para o bem da sua vida',
'positiva'),
('como se isso fosse importante', 'negativa'),
('boa tarde a todos senhores da saude e um prazer adentrar nesse grupo e quero participar 41 anos e digo vamos expandir esse programa gente o cancer mata tem jeito se nossa secretaria de saude investir melhor na qualidades de atendimentos a nos populacao carentes do brasil',
'positiva'),
('em menos de uma semana a campanha novembroazul ja entrou com tudo em campo uma goleada de informacao mobilizando os meios de comunicacao para abordar os modos de diagnosticar e prevenir o cancer de prostata a informacao correta e o nosso artilheiro nessa partida somostodosladoalado',
'negativa'),
('barra mansa nem outubro rosa tem vai ter novembro azul', 'negativa'),
('importante', 'positiva'),
('informacao sim condicao nao', 'negativa'),
('homem precisa se prevenir abracos', 'positiva'),
('sou novembro azul', 'positiva'),
('gostei', 'positiva'),
('engracado outubro rosa e novembro azul, mas diz que as cores nao tem nada haver sem querer rosa e para as mulheres e azul para os homens cade os urologistas para fazerem os exames no povo',
'negativa'),

('mensagem linda mas com galvao nao da programa nao flui so a opiniao dele que vale deixa luiz roberto', 'negativa'),
('cade o luiz roberto pra fazer o bem amigos por que aguenta o galvao so ele falando e dose ta desistindo',
'negativa'),
('nome de peso faltou o fenomeno', 'positiva'),
('na minha casa este programa quando o galvao esta nao entra', 'negativa'),
('globo lixo', 'negativa'),
('eses cara nao foro embora ainda some do brasil', 'negativa'),
('peso do que so se for em mediocridade', 'negativa'),
('globo lixo recebendo muito com isso enganando muitos com suas mediocridades de prestadora de informacao',
'negativa'),
('quando que esse lixo vai embora', 'negativa'),
('dois canalhas so eles entendem de futebol', 'negativa'),
('entao de nada adianta essa campanha', 'negativa'),
('a conscientizacao e uma parte do problema sou plenamente consciente mas dependo do sus passei em um urologista em um ame icarai o retorno era para seis meses que ja se passaram e ate agora nada uma bela campanha mas e na hora de conseguir atendimento nem nessa e na campanha passada foi abordada essa questao',
'negativa'),
('conscientizacao de uma saude melhor parabens novembro azul', 'positiva'),
('deixe o preconceito de lado', 'positiva'),
('chega novembro e ao inves de pensar na prevencao, enchem as redes sociais com muitas piadinhas que nao incentivam em nada a prevencao so aumenta o preconceito quantas pessoas ja sabem que o exame preventivo e apenas um exame de sangue simples assim se o exame apontar distorcao nos hormonios novos exames serao necessarios antes de divulgar suas piadinhas procure um clinico geral e faca uma bateria completa de exames prostata colesterol diabetes anemia entre outros cuide-se sem medo e sem preconceito',
'positiva'),
('prevencaoe necessaria', 'positiva'),
('sou da ong amigo de voce fundada em 2017 pode me dar informacoes sobre o kitpelo email axefortalezagmailcom',
'positiva'),
('muito importante o homem ter consciencia que deve cuidar da saude meu marido faz esse exame todo ano', 'positiva'),
('preconceito e o maior inimigo levou meu pai que quando viu ja era tarde', 'positiva'),
('parabens por essa acao muito importante', 'positiva'),
('grande interesse utilidade publica', 'positiva'),
('agora e vez do azul', 'positiva'),
('parabens essa acao', 'positiva'),
('queremos ver todos os homens fazendo seus preventivos contra o cancer de prostata', 'positiva'),
('pode se com o novo governo as coisa melhorar porque se nao tiver dinheiro nao tem medico nao tem exames tem locar dinheiro no balcao pra se atendido sem dinheiro nada feito estava com os exames na mao quanto nao coloquei 150 reais nao fui atendido por medico e olhando que amaioria nao tem fica dificil de fazer exames',
'negativa'),
('diagnosticar do primeiro passo quero ver o tratamento pois nao tem estrutura para isto', 'negativa'),

('nos abracamos a vida com amor', 'positiva'),
('ja aderi quando morava na inglaterra apoiava e ajudava o prostata cancer e tambem o movember e aqui o novembroazul com orgulho',
'positiva'),
('e isso ai homarada chegou a vez de voces se cuidarem tambem nos mulheres fazemos a nossa parte agora e a vez de voces ta bom a saude e a coisa mais importante das nossas vidas lembrem se disso boa sorte a todos e o que desejo beijos',
'positiva'),
('que que se previnir previne tem cura acredite se prevenir novembro azul faca os exames depois dos anos e uma obrigacao um dever e um cuidado com voce mesmo',
'positiva'),
('sempre aderir a campanha tanto das mulheres outubro rosa quanto aos homens ',
'positiva'),
('pessoal muito importante a prevencao', 'positiva'),
('uma vez por ano e importante fazer os exames de sangue e de toque', 'positiva'),
('e importante a prevencao cuide-se', 'positiva'),
('estou junto e apoiando a campanha ', 'positiva'),
('bem vindo novembro bom dia', 'positiva'),
('vamos homens inteligentes fazem o exame da prostata', 'positiva'),
('deus abencoe todo mes de novembro', 'positiva'),
('prevencao e vida', 'positiva'),
('prevencao sempre', 'positiva'),
('melhor prevenir do que remediar', 'positiva'),
('sim vamos nos cuidar', 'positiva'),
('cuidem-se', 'positiva'),
('novembroazul', 'positiva'),
('apoio', 'positiva'),
('lindos', 'positiva'),
('verdade', 'positiva'),
('juntos', 'positiva'),
('amem bom dia', 'positiva'),
('nao e bem assim, so se comenta o que deu certo mas meu caso depressao incontinencia dores etc', 'negativa'),
('vamos la pessoal fazer exames a vida e mais importante do que orgulho besta se as mulheres nao tem medo nem vergonha de fazer exames preventivos os homens tambem tem que seguir o mesmo exemplo a hora e essa nao deixe para amanha o que voce pode fazer hoje amanha podera ser tarde demais',
'positiva'),
('o mes passado foi das mulheres agora e a nossa vez e novembro azul vamos homens deixar o machismo de lado vamos cuidar da nossa saude tambem',
'positiva'),
('quem sonha viver muito reflete que beneficio trara para voce estes exames para evitar e prevenir o cancer de prostata',
'positiva'),
('sem preconceitos facam o exame deixem as mulheres orgulhosas dos homens', 'positiva'),
('vamos pessoal se cuidar pois nao devemos deixar nada para amanha para nao sofrer mais tarde', 'positiva'),
('meu maridao gutemberg moreira de barros tem anos de idade e vai fazer o exame ne amorzao amorzinho amore te love e te amo beijos da sua esposa mulher amiga nivia silva ',
'positiva'),

('vamos la falo de causa propria no comeco do ano fiz a cirurgia para retirada de 80 por cento da prostata em virtude dos exames fazia todo ano pois meu pai faleceu em virtude disso deixem os preconceitos de lado',
'positiva'),
('fiz meus primeiros exames aos anos ultra-som psa atraves do sangue exame de toque e a biopsia que nao encontrou nada comprometedor tomo um comprimido diariamente para controlar o nivel do psa faco meu tratamento no ame da cidade',
'positiva'),
('e isso homens tiro o medo do bau', 'positiva'),
('vamos la homens cuidar da saude faz bem', 'positiva'),
('quem ama e quer ser amado faz os ex preventivos saude e o lema', 'positiva'),
('todos tem que se cuidar hoje pra nao sofrer depois', 'positiva'),
('nao tenham medo', 'positiva'),
('e melhor cuidar da saude para termos', 'positiva'),
('bora se cuidar homarada', 'positiva'),
('legal', 'positiva'),
('e isso ai vamos nos cuidar beleza', 'positiva'),
('a prevencao me salvou', 'positiva'),
('vamos mais uma vez a luta', 'positiva'),
('vamos nessa ', 'positiva'),
('belissima campanha ', 'positiva'),
('faz tres meses que eu perdi meu marido com cancer de prostata', 'positiva'),
('meu pai foi diagnosticado operou em e hoje ta super bem diagnostico precoce o salvou', 'positiva'),
('salve novembro azul para os homens que deus ilumine a cabeca deles', 'positiva'),
('fazer alerta para melhorar o atendimento no sus e rapidez', 'negativa'),
('meu irmao descobriu a dois meses esta em tratamento meu outro irmao descobriu a treze anos estava bem mais novo ficou curado graca a deus este tambem esta entregue na suas maos',
'positiva'),
('vamos la homens marido filho irmaos sobrinhos primos genros cunhados todos', 'positiva'),
('ai homens de todo o mundo todos voces rosa foi mulheres agora chegou vez de voces cuidar da saude vamos la em todos os do mundo cuide se a hora e esta fica dia domingo azul pra voces valeu',
'positiva'),
('parabens olha ai homens se cuide e muito importante pra todos fazer o exame pra se prevenir contra essa doenca tao terrivel',
'positiva'),
('a vida nao e um jogo ', 'positiva'),
('fiquem atentos homens o mes de novembro e pra alertar se cuidem olha o faustao alertando', 'positiva'),
('fica a dica homens', 'positiva'),
('e todos os homens do mundo principalmente aqueles com mais de 40 e muito importante pra voces ', 'positiva'),
('legal fausto e isso ai tem que se cuidar', 'positiva'),
('novembro azul eu apoio prevencao a melhor opcao amigos', 'positiva'),
('quem ama se cuida', 'positiva'),
('vamos incentivar aos homens para se cuidarem', 'positiva'),
('essa pra os homens se cuidar', 'positiva'),

('e isso ai', 'positiva'),
('parabens', 'positiva'),
('faustao e saude', 'positiva'),
('e novembro azul para homens', 'positiva'),
('bora se prevenir meus amigos cidinho gregorio jose aparecido entre outros', 'positiva'),
('fiquem atentos prevenir e a melhor solucao pense nisso deus abencoe voces', 'positiva'),
('adorei a prevencao ainda e a melhor coisa', 'positiva'),
('prevencao sim', 'positiva'),
('parabens homem que se cuida', 'positiva'),
('nao faco preventivo sigo a dita dos especialistas dizem eles que os acertos, tanto do toque retal quanto do psa sao da ordem de a por cento ha portanto um erro de a por cento sendo que mil cirurgias foram realizadas sem necessidade fazendo com que esse universo de homens impotentes e com incontinencia urinaria passem a conviver infeliz dizem tambem que se houver o cancer somente um ano de vida aconselham esquecer o problema levar uma vida regrada alimentacao equilibrada e se o jato da miccao for forte va viver a vida e esqueca a prostata',
'negativa'),
('nada a ver so propaganda nao funciona a menos se for convenio porque pelo sus estamos perdidos a media mostra quando da certo o meu caso ficou sequelas irreversiveis depressao dores e o inss ainda humilhando',
'negativa'),
('sera que dessa vez vai acontecer ou sera so palestras preventivas teremos exames ', 'negativa'),
('corrijam a data postada nao seria quarta-feira', 'negativa'),
('e isso ai', 'negativa'),
('sim e melhor assim', 'positiva'),
('vamos aniquilar o cancer fosfeotomina gostaria de saber mais desse remedio que a usp descobriu a muito tempo contra o cancer',
'positiva'),
('homens deixem seu medico de confianca fazer o exame e boa sorte', 'positiva'),
('valeu e isso ai', 'positiva'),
('parabens bonito gesto', 'positiva'),
('vamos homens valorosos coragem', 'positiva'),
('por favor, eu sempre fiz e acabei descobrindo cedo o que me permitiu tratar mas nao escapei da cirurgia',
'positiva'),
('legal', 'positiva'),
('temos que ter coragem fazer o exame toque retal garantia', 'positiva'),
('como fazer em postinho de saude de prefeitura', 'positiva'),
('prevencao para todos', 'positiva'),
('por favor olhe que estao divulgando sim com antecedencia pela causa', 'positiva'),
('coragem homens prevencao', 'positiva'),
('vamos sim participar temos que ser homem mesmo', 'positiva'),
('gente preciso de depoimento para tv, de pessoas que tiveram cancer de prostata ou estao passando pelo processo de tratamento alguem me chamem no inbox por favor novembroazul',
'positiva')]

Base de testes

('muita boa sorte a todos os homens deus esteja sempre presente na vida de todos voces bom dia a todos', 'positiva'),
('agora e a vez de voces pai filho irmaos tios genros cunhados pessoal vamos la se cuidar', 'positiva'),
('verdade eu hoje ja estou passando no medico para fazer diversos exames anuais', 'positiva'),
('que todos os homens se conscientizem da importancia da prevencao', 'positiva'),
('boa sorte a todos os homens deus abencoe a todos', 'positiva'),
('estamos juntos', 'positiva'),
('muito bom', 'positiva'),
('somostodosladoalado', 'positiva'),
('novembro azul estamos nessa', 'positiva'),
('ta certo hugo vamos la garoto', 'positiva'),
('sensacional', 'positiva'),
('apoiado', 'positiva'),
('nao precisa esperar novembro deixe o medico tocar sua prostata ao menos uma vez por ano assim como nao precisamos esperar outubro para fazer mamografia', 'positiva'),
('quanta baboseira vai no postinho pra consultar so pra dezembro ai ja acabou a campanha', 'negativa'),
('tamo junto nessa', 'positiva'),
('sensacional', 'positiva'),
('as disfuncoes sexuais masculinas e femininas podem afetar os relacionamentos afetivos entre os homens a ejaculacao precoce e a disfuncao eretil sao os problemas mais comuns vencer o tabu e conversar sobre o assunto e o primeiro passo para supera-lo converse com sua parceira(o) a respeito e busque ajuda profissional', 'positiva'),
('a prevencao do cancer de prostata deve atraves do exame ultra som psa no sangue ou toque retal e preciso fazer os exames preventivos', 'positiva'),
('novembro azul prevenir melhor que remediar facam o alto exame psa toque nao deixe que o preconceito traga prejuizo para sua saude', 'positiva'),
('a saude do homem sem preconceitos', 'positiva'),
('novembro azul', 'positiva'),
('seja homem seja feliz faca sua parte faca prevencao', 'positiva'),
('prevencao do cancer de prostata', 'positiva'),
('estamos junto', 'positiva'),
('estamos juntos', 'positiva'),
('otima lembranca', 'positiva'),
('meus filhos e eu claro', 'positiva'),
('ta devagar', 'positiva'),
('bora', 'positiva'),
('conhecer as propostas de politicas publicas para a saude dos cidadaos e um direito seu e um dever de toda a sociedade informe-se sobre a plataforma de cada candidato e sobre o compromisso com a sua saude', 'positiva'),
('a saude masculina tem um impacto direto na saude da mulher cuidar do seu corpo e respeitar sua parceira e construir uma relacao saudavel e igualitaria voce tem um filho adolescente converse abertamente com ele sobre saude sexual e cuidados com o parceiro a vacinacao de meninos entre onze e quinze anos contra o hpv pode ajudar a diminuir os riscos de cancer genital nos homens e cancer de colo de utero nas mulheres', 'positiva'),
('verdade hoje os pais tem essa abertura o que antes era tabu hoje e prevencao', 'positiva'),

('nos mulheres temos que nos lembrar de apoiar o filho neste sentido', 'positiva'),
('precisamos falar de prevencao e saude', 'positiva'),
('cuide-se', 'positiva'),
('eu apoio essa ideia', 'positiva'),
('ue nao estamos em outubro', 'negativa'),
('amei a foto do perfil rosamas seria o caso divulgar outubro azul(cor azul)', 'negativa'),
('uaiiii nao tem nada azul na foto', 'negativa'),
('atitudes como essa devem ser exaltadas parabens aos organizadores', 'positiva'),
('voce pode nao saber mas os urologistas se dedicam a cuidar da saude e bem-estar de todos nos eles tratam da saude masculina de modo integrado com outras especialidades garantindo mais qualidade de vida para o homem e trazendo impactos positivas para a saude de toda a familia', 'positiva'),
('parabens em vez de novembro azul faca como eu para nao esquecer tenha muitos aniversarios azuis parabens', 'positiva'),
('heteros bissexuais e gays todos somos parceiros saude sempre', 'positiva'),
('muito importante', 'positiva'),
('que bom', 'positiva'),
('utopia isso nao sus mal atende o pre natal feminino imagine saude masculina', 'negativa'),
('importante', 'positiva'),
('muito bem', 'positiva'),
('eu adorei e tambem curti', 'positiva'),
('gostei muito curti e convenio', 'positiva'),
('assim que deve ser continuadada e faz o tratamento de graca quer dizer de graca nao esqueci nesse pais tem alguma coisa de graca se pagamos tudo pagamos caro e pouco temos no retorno quando temos deus na causa ', 'negativa'),
('muitos homens evitam procurar ajuda medica para resolver problemas de carater sexual e reprodutivo muitos deles facilmente trataveis mas quanto mais cedo assumirem que podem precisar de apoio e aconselhamento medico mais qualidade de vida ganham repense suas escolhas', 'positiva'),
('eu me cuido', 'positiva'),
('com todos estes sintomas procurar um medico no sus para ser atendido daqui seis meses', 'negativa'),
('o que fazer com a nossa saude so basta rezar', 'negativa'),
('se estao acabando com aposentadoria imagina so a saude sugestao poderia fazer na periferia', 'negativa'),
('radio rainha da paz e bem estar pra voce se coisa louvavel revelar e publicar as obras do senhor nossa senhora das gracas rogai por nos', 'positiva'),
('muito bom o alerta mas devem ser divulgados as outras do cancer de penis', 'positiva'),
('nos cuidemos sempre', 'positiva'),
('obrigado pela iniciativa precisamos mais desse tipo de informacao', 'positiva'),
('muitos colegas acham que e brincadeira mas e muito serio essa questao', 'positiva'),
('vai ter medicos exames remedios se tiver otimo que nao seja mais uma ilusao', 'positiva'),
('se isso fosse verdade mais tambem pra uma pessoa pobre e ao medico e na farmacia basica ate remedio ai adianta so se tomar cha da receita', 'negativa'),
('muitas boas intencoes aguardamos', 'positiva'),
('acho bom mesmo que funcione pois quando chega o horario da meia noite voce vai na ubs do bairro santa ines e so tem o vigilante trancado e o povo fica a madrugada na fila

esperando que amanheca pro posto abrir um absurdo isso prefeito clecio por favor faca uma vistoria pois eu mesma ja vi essa situacao', 'negativa'),

('brasil rumo ao sistema humanizado sus melhor sistema de saude implantado para servir aos brasileiros', 'positiva'),

('se tiver medicamentos pessoal bem remunerado pode ser mas tenho duvidas', 'negativa'),

('cortinas de fumaca a saude e mais que atendimento ambulatorial numeros nao significam qualidade e sem qualidade nao existem resultados efetivos uma nova saude publica e preciso', 'negativa'),

('em barra mansa nada funciona faz cinco anos que estou tentando fazer mamografia e ate hoje nada', 'negativa'),

('do que adianta se nao ha medicos no local pra da um atendimento humano sem falar da falta de medicamentos', 'positiva'),

('so nao acho certo abrir concurso publico para varios cargos e deixarem os tecnicos de enfermagem a deus dara tdos entram na saude publica e nos tecnicos de enfermagem temos que ajudar a pagar os funcionarios que nao faz parte da equipe(psf)', 'negativa'),

('prometem mais nao cumprem', 'positiva'),

('parabens', 'positiva'),

('o descompaco entre as leis e a execucao dos servicos publicos ocorre principalmente pela falta de compromisso do poder publico uma vez que os recursos financeiros repassados sao insuficientes para garantir equipamentos insumos servicos de exames de laboratorios e de imagem e recursos humanos suficientes e qualificados para atender a grande demanda ou seja o numero enorme de pessoas com os diversos tipos de cancer e o atual governo reduz os gastos e faz corte significativo na saude entao a tendencia e piorar a saude no pais e uma estrategia do governo atual para privatizar a saude precisamos nos organizar e lutar por um sus publico universal e de qualidade', 'positiva'),

('os clientes tambem nao podia esperar mais do que 15 minutos na fila do banco hoje estao esperando so umas 4 horas pra serem atendidos lixo lixo lixo de pais', 'negativa'),

('nao adianta gente nada acontece implore pela vida a deus deus e mais aqui e brasil um lixo de pais', 'negativa'),

('faco todos os anos o pse e a tira de prostata', 'positiva'),

('vamos nos cuidar', 'positiva'),

('isso tem de falar para o governo pois tem pessoas esperando pra fazer um exame a mais de dois anos', 'negativa')]