Um modelo computacional para identificação de notícias falsas sobre a Covid-19 no Brasil

Débora da Conceição Araújo
debora.caraujo@univasf.edu.br

Resumo. Em dezembro de 2019 apareceram na China os primeiros casos de humanos infectados por um novo coronavírus, o Sars-Cov-2. O vírus apresentou um alto poder de contágio e em maio de 2020 já haviam registros da sua presença em mais de 200 países e mais de 300 mil mortes em decorrência da Covid-19, doença causada pelo Sars-Cov-2. Ainda não há vacinas ou remédios contra o vírus disponíveis à população. Imersa em um cenário de incertezas, a população iniciou um fenômeno de propagação de notícias falsas, do inglês *fake news*, a respeito do novo coronavírus. Tais notícias estão relacionadas à origem do Sars-Cov-2, aos sintomas, formas de contaminação e tratamento da Covid-19. Dentro desse contexto, este projeto tem por objetivo auxiliar no combate a propagação de *fake news* sobre o novo coronavírus e propõe a construção de um modelo computacional capaz de identificar este tipo de notícia nas redes sociais. Para avaliação do modelo, será construída uma base de dados a partir de postagens do Twitter com informações compartilhadas sobre a Covid-19, considerando a geolocalização do Brasil, país epicentro da doença na América do Sul. Por meio do método proposto será possível notificar sites e usuários que estão em contato com as notícias falsas de forma rápida e automática, de modo a minimizar os impactos da desinformação sobre a o novo coronavírus no Brasil.

1. Introdução

Pandemia é a rápida disseminação de uma doença e ocorre, em geral, quando um vírus se espalha entre países e continentes em um curto espaço de tempo, como dias ou meses (CUNHA, 2004). Aspectos como gravidade e índice de mortalidade não são determinantes para se caracterizar um cenário de pandemia, para tal são avaliados o poder de contágio e a capacidade de proliferação geográfica da doença. De acordo com a Organização Mundial da Saúde (OMS, 2011), a transmissão simultânea de um vírus da gripe, em nível mundial, é suficiente para se indicar uma pandemia.

No ano de 2009, um cenário recente de pandemia foi causado pela influenza A (H1N1), infecção respiratória que apresentou um baixo índice de letalidade, variando entre 0,01% e 0,03% (WHO, 2011), mas um alto potencial de contágio. Em menos de um ano, haviam casos registrados da influenza em mais de 200 países (WHO, 2010). De forma mais agressiva do que a influenza A (H1N1), o novo coronavírus, o Sars-Cov-2, ultrapassou os 200 países com registros de infectados em menos de 6 meses após sua descoberta. Os primeiros registros surgiram em dezembro de 2019, quando apareceram 27 casos considerados, inicialmente, como pneumonia, na cidade Wuhan, China (KANG; XU, 2020).

O vírus Sars-Cov-2 tem os morcegos como hospedeiros naturais e ainda não existem estudos conclusivos sobre sua migração para o organismo humano. Quando uma pessoa já está infectada, a transmissão do vírus entre humanos ocorre por meio de gotículas eliminadas por saliva, como tosse ou espirros. A contaminação ocorre também quando a pessoa entra em contato com superfícies infectadas, seja por contato direto, como um aperto de mãos, ou por contato indireto, como o toque em demais superfícies contaminadas (WHO, 2020).

Covid-19 é o nome científico da doença causada pelo Sars-Cov-2. A doença afeta as pessoas de maneiras muito distintas: algumas são assintomáticas e não manifestam nenhum tipo de efeito; outras apresentam sintomas como tosse seca, febre e cansaço (WHO, 2020). Efeitos menos comuns como diarréia, dor de garganta e cabeça, conjuntivite, perda de paladar ou olfato e erupções cultâneas na pele também podem ser apresentados. Entre os sintomas mais graves da Covid-19 estão dificuldades com a respiração e falta de ar, dor ou pressão no peito e perda de fala ou movimentos (WHO, 2020).

O novo coronavírus se espalhou de forma rápida entre os continentes. Em março de 2020 a quantidade de casos da doença na Itália ultrapassou a quantidade de casos registrados na China e o país se tornou o epicentro da Covid-19 na Europa. Desde então, países como Espanha, Rússa e Reino Unido também ultrapassaram o número de registros chinês da Covid-19. Na América do Norte, os Estados Unidos aparecem como epicentro da doença, registrando mais de um milhão e meio de casos em maio de 2020. Na América do Sul, o Brasil é o país com maior proliferação do vírus, ultrapassando os 300 mil casos e 20 mil mortos (BRASIL, 2020a). Em todo o mundo são mais de 5 milhões de casos confirmados da Covid-19, além de mais de 325 mil mortes (EU open data, 2020).

Com a proliferação do Sars-CoV-2 na América Latina, observou-se uma crescente propagação de notícias falsas, do inglês *fake news*, relacionadas à doença nas Redes Sociais (RS). A rede social Twitter, com o intuito de conter a disseminação desses conteúdos em sua plataforma, vêm removendo postagens que espalham desinformação sobre a Covid-19, inclusive desinformações anunciadas por chefes de Estado (LYONS, 2020).

No Brasil, epicentro da doença na América do Sul, o combate as fake news motivou a proposição de um projeto que cria a Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet. De acordo com o Projeto de Lei (PL), considera-se como desinformação ou notícia falsa o conteúdo enganoso que foi colocado fora de contexto, manipulado ou completamente forjado com o interesse de enganar a população (BRASIL, 2020b). O PL prevê que quando um conteúdo com alcance significativo for considerado desinformação, as RS devem implementar medidas para minimizar a disseminação do conteúdo. Entende-se que as redes sociais possuem papel fundamental para o combate às fake news na internet. Com algoritmos que consideram as preferências do usuário, as pessoas têm muito contato com conteúdos que corroboram com sua visão de mundo, o

que não significa que há evidências científicas que comprovem as informações compartilhadas (CINELLI et al., 2020).

Ao considerar o contexto delineado, este estudo objetiva propor um modelo computacional capaz de identificar, de forma automática, notícias falsas sobre a Covid-19 no Brasil. Para tal, serão analisadas postagens presentes na rede social Twitter. A escolha do Twitter para análise ocorre, pois cerca de 41 milhões de brasileiros fazem uso desta plataforma; além disso, a influência das notícias falsas compartilhadas nesta RS é alvo de estudo de diversos trabalhos presentes na literatura (BUNTAIN;GOLBECK, 2017; AJAO; BHOWMIK; ZARGARI, 2018; BOVET; MAKSE, 2019; CINELLI et al., 2020).

As principais contribuições deste projeto podem ser sumarizadas da seguinte forma:

- Construção e rotulação de uma base de dados com informações compartilhadas no Twitter sobre a Covid-19, considerando o idioma português e a localização geográfica do Brasil;
- Análise da propagação de notícias falsas sobre a Covid-19 no Brasil, por Regiões e Estados;
- Levantamento sobre o teor das notícias falsas mais compartilhadas sobre a Covid-19;
- Construção e análise de um modelo computacional capaz de identificar notícias falsas sobre a Covid-19 no idioma português.

2. Justificativa

A disseminação de fake news é um problema nos diversos âmbitos da sociedade e tem motivado a condução de inúmeras pesquisas no mundo (ZHANG; GHORBANI, 2020). Tais notícias são enquadradas em três tipos principais: Serious Fabrications (reportagens fraudulentas), Large-Scale Hoaxes (farsas em larga escala) e Humorous Fakes (notícias humorísticas) (RUBIN et al., 2015), sendo:

- Reportagens fraudulentas – tentativas de enganar o público com notícias embaraçosas, presentes na mídia convencional;
- Farsas em larga escala - notícias modificadas ou criadas com o intuito de espalhar desinformação;
- Humorísticas - notícias humorísticas que se utilizam de sarcasmo e ironia. É importante que os leitores estejam cientes da intenção humorística da notícia.

De acordo com Silva et al. (2020), a maior parte das fake news se enquadra no perfil de Farsas em larga escala. Esse tipo de notícia ganhou notoriedade após as eleições presidenciais de 2016 nos Estados Unidos, quando houveram mais de 8 milhões de comentários, reações e compartilhamentos na RS Facebook, envolvendo notícias falsas sobre as eleições e seus candidatos (HORNE; ADALI, 2017; ZHOU et al., 2019; ZHANG; GHORBANI, 2020).

Em Waszak et al. (2018) é avaliada a disseminação de notícias falsas voltadas à saúde. Foram analisados os links relacionados à saúde mais compartilhados entre os anos de 2012 e 2017, no idioma polonês. Cada link foi verificado quanto à presença de notícias falsas. Os resultados mostraram que 40% dos links compartilhados com mais frequência continham textos enganosos. Os links com conteúdo enganoso foram compartilhados mais de 450.000 vezes e "vacinas" era o assunto mais recorrente.

Pesquisas voltadas à saúde também são apresentadas em Lavorgna et al. (2018), em que foram analisadas as postagens de uma comunidade italiana, supervisionada por médicos, dedicada a informações sobre esclerose múltipla. Foram verificadas as postagens de usuários identificados como influenciadores e, dentre as 380 postagens com informações médicas, 72 apresentaram desinformação.

Para detecção automática de notícias falsas envolvendo as diversas áreas podem ser utilizados diferentes classificadores de Aprendizado de Máquina (AM). Algoritmos de AM identificam padrões a fim de inferir conhecimento a partir de um conjunto de dados de treinamento. O aprendizado é testado por meio de outro conjunto com novos dados, no qual o algoritmo deve classificar corretamente os objetos (MICHIE et al., 1994).

Em Skeppstedt et al. (2017) o classificador Support Vector Machine (SVM) foi utilizado, junto a técnicas de pré-processamento de texto, para analisar comentários sobre vacinas. Para tal, foram avaliadas as postagens de participantes de seis fóruns de discussão, presentes no site britânico Mumsnet. A métrica f-score (FERRI et al., 2009) foi utilizada para avaliar o desempenho do SVM. O modelo alcançou uma acertividade de 62% para a classificação binária que considera os rótulos contra ou a favor da vacinação.

Em Reis et al. (2019) são utilizados os algoritmos K-Nearest Neighbors (KNN), Naive Bayes (NB), Random Forest (RF), SVM e XGBoost (XGB) com o intuito de detectar fake news acerca das eleições presidenciais dos EUA em 2016. A f-score foi utilizada para verificar o desempenho dos classificadores e os algoritmos RF e XGB obtiveram os melhores desempenhos médios, 81% de f-score.

Este projeto se destaca aos demais por propor uma abordagem que considera o contexto de propagação de notícias falsas acerca da pandemia de Covid-19. Entre as principais contribuições estão a construção e rotulação de uma base de dados sobre o coronavírus e a Covid-19 para o idioma português, considerando a geolocalização do Brasil. Serão analisados os desempenhos de algoritmos de AM para as tarefas de classificação de texto e, por fim, uma combinação de classificadores será proposta para análise.

3. Objetivos

Objetivo Geral. Propor um modelo computacional capaz de identificar notícias falsas sobre a Covid-19 em português.

Objetivo específico 1. Construir um crawler/robô que selecione tuítes (postagens) relacionados a Covid-19, considerando o idioma português e a localização geográfica do Brasil;

Objetivo específico 2. Rotular a base de dados de acordo com as classes news ou fake news;

Objetivo específico 3. Analisar o teor das notícias falsas mais compartilhadas, além das características de sua disseminação, considerando a propagação por Regiões e Estados brasileiros;

Objetivo específico 4. Selecionar os modelos de Aprendizado de Máquina que serão combinados para a tarefa de identificação das notícias falsas;

Objetivo específico 5. Avaliar o desempenho dos modelos construídos.

4. Metodologia

4.1 Base de Dados

A base de dados utilizada neste projeto será construída a partir de postagens (tuítes) realizadas na rede social Twitter. Uma combinação (string) de busca será definida por meio de pesquisa sobre o novo coronavírus e sobre a Covid-19. Para que um tuíte seja selecionado é necessário que esteja marcado com a privacidade 'pública', além de atender aos requisitos da string de busca definida. Para realizar a seleção de tuítes de modo automático, será construído um crawler/robô que seleciona as postagens levando em consideração a privacidade do tuíte, a string de busca definida, o idioma e a geolocalização de origem da postagem.

Com a base de dados montada, inicia-se a etapa de rotulação, em que o pesquisador irá analisar as postagens e identificar a quantidade de classes (rótulos) existentes na base. A hipótese inicial deste projeto sugere que existam dois rótulos: news (notícias verdadeiras) e fake news (notícias falsas). Uma terceira classe poderá ser adicionada após análise da base. Definidas as classes, o pesquisador deverá rotular a base de dados manualmente, para que os rótulos sejam utilizados no treinamento e teste/avaliação dos modelos de AM que serão construídos. Os objetivos 1, 2 e 3 serão atingidos nesta etapa.

4.2 Pré-processamento dos dados

De acordo com Chen et al (2018), quando se trata de aprendizado de máquina, o pré-processamento dos dados é um procedimento-chave para o desempenho dos métodos de classificação. As técnicas utilizadas na etapa de pré-processamento de textos podem variar de acordo com as características de cada base de dados. Neste projeto serão utilizadas as seguintes técnicas:

- remorção de stopwords - consiste na exclusão de palavras que não contribuem para o significado mais profundo do texto. A remoção de stopwords atinge, principalmente, palavras que representam artigos, conjunções e preposições (LO; HE; OUNIS, 2005);
- Stemming - método que consiste em reduzir as palavras ao seu radical (PLINSSON; LAVRAC; MLADENIĆ, 2004). Esse método pode beneficiar a classificação do texto, tanto por reduzir o vocabulário de palavras quanto por se concentrar no sentido completo do texto, em vez de analisar o significado de cada palavra de modo individual;
- Vetorização/Discretização - modelos de AM não trabalham com textos em linguagem natural, portanto existem técnicas que visam associar os elementos textuais a valores discretos. Neste projeto será utilizada a técnica Term-Frequency Inverse Document Frequency (TFIDF), que possui eficácia comprovada em diversos trabalhos presentes na literatura (ZHU et al., 2016; GUPTA et al., 2018).

Além das técnicas supracitadas, outros métodos de mineração de textos poderão ser utilizados, de acordo com as características observadas na base de dados.

4.3 Modelos de Aprendizado de Máquina

Em AM o aprendizado pode acontecer de forma supervisionada, com dados previamente rotulados; não-supervisionada, para dados não-rotulados e; semi-supervisionada, com um pequeno conjunto de dados rotulados e a maior parte do conjunto sendo não-rotulada.

Este projeto propõe a utilização de algoritmos que fazem uso do aprendizado supervisionado, tendo em vista a etapa de construção e rotulação da base definida. Os algoritmos selecionados para realização dos teste iniciais são:

- Rede Neural Artificial – baseada em um sistema nervoso, seu funcionamento conta com a presença de neurônios artificiais interligados através de sinapses (pesos). É realizado um somatório ponderado no núcleo do neurônio, entre as entradas e seus respectivos pesos, e com base em um limiar de ativação é verificado se a entrada será ou não propagada entre os neurônios das camadas adjacentes até a camada de saída (WANKHEDE, 2014).
- Naive Bayes - baseado no Teorema de Bayes, é conhecido como um classificador simples, entretando, apresenta resultados satisfatórios em diversos problemas do mundo real (RISH, 2001). Esse algoritmo calcula a probabilidade de um dado elemento pertencer a uma determinada classe/rótulo.
- Support Vector Machine - baseado na teoria do aprendizado estatístico, foi pensado para problemas lineares, mas também consegue lidar com problemas não-lineares. O algoritmo objetiva encontrar um hiperplano que divida os rótulos em um plano cartesiano. Este hiperplano é obtido a partir do treinamento dos elementos para, posteriormente, classificar os dados de teste (VAPNICK, 2013).

Testes com outros algoritmos poderão ser realizados, caso se observe necessário ao decorrer da pesquisa. Ao final, será proposto, a partir dos algoritmos utilizados, um comitê de classificadores para a detecção de notícias falsas sobre a Covid-19. O método de comitê parte da premissa de que a combinação de vários modelos, treinados separadamente, pode aumentar significativamente a capacidade de generalização de um sistema (LIMA, 2017). Por fim, o comitê proposto terá seu desempenho comparado aos demais algoritmos utilizados para testes.

Para viabilizar as análises estatísticas e comparação entre os desempenhos dos modelos, será utilizado o método *k-fold crossvalidation* (BROWNE, 2000). Com esse método, serão realizadas *k* repetições de cada classificador, de modo a se atingir uma amostra de tamanho *y* com os resultados dos modelos. A partir dessas amostras torna-se possível analisar se há diferença estatisticamente significante entre os desempenhos dos diferentes classificadores.

As análises dos desempenhos dos modelos devem iniciar com um teste de normalidade dos dados, como o Shapiro-Wilk (Ghasemil; Zahediasl, 2012). Caso as amostras obedeçam a uma distribuição normal, o teste paramétrico t-Student (SILVA, 2017) poderá ser utilizado, caso não obedeçam a uma distribuição normal, uma alternativa não paramétrica será utilizada, como o teste de Wilcoxon (SILVA, 2017).

5. Produção científica e Contribuições esperadas

Com a realização deste projeto pretende-se contribuir com as pesquisas de Computação Aplicada à Saúde por meio dos seguintes pontos:

- Combate a propagação de notícias falsas sobre a área da saúde, em especial sobre o Sars-Cov-2 e a Covid-19;
- Identificação do perfil do brasileiro que está suscetível a contribuir com a disseminação de notícias falsas;
- Construção de um modelo computacional capaz de identificar postagens que contenham notícias falsas sobre a Covid-19 e sobre o Sars-Cov-2.
- Com o desenvolvimento do projeto, artigos poderão ser submetidos a periódicos como *Expert Systems with Applications* e *Computers in Human Behavior*.

Literatura Citada

AJAO, Oluwaseun; BHOWMIK, Deepayan; ZARGARI, Shahrzad. Fake news identification on twitter with hybrid cnn and rnn models. In: Proceedings of the 9th International Conference on Social Media and Society. p. 226-230, 2018.

BOVET, A.; MAKSE, Hernán A. Influence of fake news in Twitter during the 2016 US presidential election. Nature communications, v. 10, n. 1, p. 1-14, 2019.

BRASIL. Ministério da Saúde. Painel Coronavírus. Brasília, DF, 2020a.

BRASIL. Projeto de Lei nº 1429/2020, de 01 de abril de 2020. Institui a Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet. Brasília: Assembleia Legislativa, [2020]. Disponível em: <https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2242713>. Acesso em: 21 maio 2020, 2020b.

BROWNE, Michael W. Cross-validation methods. Journal of mathematical psychology, v. 44, n. 1, p. 108-132, 2000.

BUNTAIN, Cody; GOLBECK, Jennifer. Automatically identifying fake news in popular Twitter threads. In: IEEE International Conference on Smart Cloud (SmartCloud). IEEE, 2017. p. 208-215, 2017.

CHEN, Yong et al. Sentiment Analysis Based on Deep Learning and Its Application in Screening for Perinatal Depression. In: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC). IEEE, 2018. p. 451-456, 2018.

CINELLI, Matteo et al. The covid-19 social media infodemic. arXiv preprint arXiv:2003.05004, 2020.

CUNHA, B. A. Influenza: historical aspects of epidemics and pandemics. Infectious Disease Clinics, 18(1), 141-155, 2004.

European Union open data. EU open data, 2020. COVID-19 Coronavirus data. atual. 18 maio 2020. Disponível em: <https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data>. Acesso em: 20 maio 2020.

FERRI, César; HERNÁNDEZ-ORALLO, José; MODROIU, R. An experimental comparison of performance measures for classification. Pattern Recognition Letters, v. 30, n. 1, p. 27-38, 2009.

GHASEMI, Asghar; ZAHEDIASL, Saleh. Normality tests for statistical analysis: a guide for non-statisticians. International journal of endocrinology and metabolism, v. 10, n. 2, p. 486, 2012.

GUPTA, Mehul et al. A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers. In: 2018 Eleventh International Conference on Contemporary Computing (IC3). IEEE, 2018. p. 1-7, 2018.

HORNE, B. D.; ADALI, Sibel. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: Eleventh International AAAI Conference on Web and Social Media. 2017.

JOEL, Plisson; LAVRAC, Nada; MLADENIC, Dunja. A rule based approach to word lemmatization. In: Proceedings C of the 7th International Multi-Conference Information Society IS. 2004.

KANG, Y.; XU, S. Comprehensive overview of COVID-19 based on current evidence. Dermatologic Therapy, 2020.

LAVORGNA, L. et al. Fake news, influencers and health-related professional participation on the Web: A pilot study on a social-network of people with Multiple Sclerosis. Multiple sclerosis and related disorders, v. 25, p. 175-178, 2018.

LIMA, T. P. F. d. Sistema híbrido inteligente para geração, seleção e combinação de classificadores. Tese (Doutorado) — Universidade Federal de Pernambuco, 2017.

LO, R. Tsz-Wai; HE, Ben; OUNIS, Iadh. Automatically building a stopword list for an information retrieval system. In: Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR). p. 17-24, 2005.

Lyons, K. Twitter removes tweets by Brazil, Venezuela presidents for violating COVID-19 content rules. The Verge, Nova Iorque, 30 Mar. 2020. Disponível em: <https://www.theverge.com/2020/3/30/21199845/twitter-tweets-brazil-venezuela-presidents-covid-19-coronavirus-jair-bolsonaro-maduro>. Acesso em: 20 maio, 2020.

MICHIE, Donald et al. Machine learning. Neural and Statistical Classification, v. 13, n. 1994, p. 1-298, 1994.

REIS, J.C.S. et al. Supervised learning for fake news detection. IEEE Intelligent Systems, v. 34, n. 2, p. 76-81, 2019.

RISH, Irina et al. An empirical study of the naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. 2001. p. 41-46.

RUBIN, V. L.; CHEN, Yimin; CONROY, N. J. Deception detection for news: three types of fakes. Proceedings of the Association for Information Science and Technology, v. 52, n. 1, p. 1-4, 2015.

SILVA, E. G. d. Previsão de séries temporais usando sistemas de múltiplos preditores. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2017.

SILVA, R. M. et al. Towards Automatically Filtering Fake News in Portuguese. Expert Systems with Applications, p. 113199, 2020.

SKEPPSTEDT, Maria; KERREN, Andreas; STEDE, Manfred. Automatic detection of stance towards vaccination in online discussion forums. In: Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017). p. 1-8, 2017.

VAPNIK, Vladimir. The nature of statistical learning theory. Springer science & business media, 2013.

WANKHEDE, Sonali B. Analytical study of neural network techniques: SOM, MLP and classifier-a survey. IOSR Journal of Computer Engineering, v. 16, n. 3, p. 86-92, 2014.

WASZAK, Przemyslaw M.; KASPRZYCKA-WASZAK, Wioleta; KUBANEK, Alicja. The spread of medical fake news in social media–the pilot quantitative study. Health policy and technology, v. 7, n. 2, p. 115-118, 2018.

WORLD Health Organization. WHO, 2010. Pandemic (H1N1) 2009 - update 100. Disponível em: <https://www.who.int/csr/disease/swineflu/laboratory14_05_2010/en/>. Acesso em: 20 maio 2020.

WORLD Health Organization. WHO, 2011. The classical definition of a pandemic is not elusive. Disponível em: <https://www.who.int/bulletin/volumes/89/7/11-088815/en/>. Acesso em: 20 maio 2020.

WORLD Health Organization. WHO, 2020. Coronavirus. Disponível em: <https://www.who.int/health-topics/coronavirus#tab=tab_3>. Acesso em: 20 maio 2020.

ZHANG, Xichen; GHORBANI, A. A. An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, v. 57, n. 2, p. 102025, 2020.

ZHOU, Xinyi et al. Fake news: Fundamental theories, detection strategies and challenges. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, p. 836-837, 2019.

ZHU, Wei et al. A study of damp-heat syndrome classification using Word2vec and TF-IDF. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2016. p. 1415-1420.