Tendências sobre saúde nas mídias sociais: Análise usando a modelagem de tópicos do Twitter

Mohsen Asghari
Departamento de Engenharia da Computação e Ciência da Computação
CECS University of Louisville
Louisville, KY, EUA
m0asgh02@louisville.edu

Daniel Sierra-Sosa
Departamento de Engenharia da Computação e Ciência da Computação
CECS University of Louisville
Louisville, KY, EUA
d.sierrasosa@louisville.edu

Adel Elmaghraby
Departamento de Engenharia da Computação e Ciência da Computação
CECS University of Louisville
Louisville, KY, EUA
adel@louisville.edu

Resumo—Há um interesse crescente nas redes sociais por temas relacionados à Saúde. Em particular, no Twitter, milhões de tweets relacionados à saúde podem ser encontrados. Esses posts contêm opiniões públicas sobre saúde e permitem entender como é a percepção popular sobre temas como diagnóstico médico, medicamentos, instalações e reivindicações. Neste artigo apresentamos um sistema adaptativo projetado usando 5 camadas. O sistema contém uma combinação de algoritmos não supervisionados e supervisionados para rastrear as tendências das mídias sociais de saúde. Como é baseado em um modelo word2vec, ele também captura a correlação de palavras com base no contexto, melhorando ao longo do tempo, aprimorando a precisão das previsões e rastreamento de tweets. Neste trabalho, focamos nos dados dos Estados Unidos e os usamos para detectar os trending topics de cada estado. Esses tópicos são acompanhados incluindo novas contribuições nas redes sociais. O algoritmo supervisionado implementado é uma Rede Neural Convolucional (CNN) em conjunto com o modelo Word2Vect para classificar e rotular novos tweets, atribuindo um feedback aos modelos de tópicos. Os resultados deste algoritmo apresentam acurácia de 83,34%, precisão de 83%, recall de 84% e F-Score de 83,8% quando avaliados. Nossos resultados são comparados com duas técnicas de última geração, demonstrando uma vantagem que pode ser aproveitada para melhorias adicionais.

Palavras-chave: tweets de saúde, LDA, classificação, aprendizado profundo

I. INTRODUÇÃO

O texto livre na área da saúde é classificado em dois grupos, o texto biomédico e o texto clínico. O texto biomédico inclui livros, resumos e artigos. O texto clínico compreende relatórios do pessoal médico, como as patologias do paciente diagnosticadas, histórico pessoal e médico [1]. No entanto, textos relacionados à saúde também estão disponíveis na rede social como livres/não estruturados texto constituindo mais um grupo de texto.

Sites de redes sociais como Twitter ou Facebook fornecem uma plataforma de comunicação. Estudos recentes mostram que no Twitter os usuários tendem a compartilhar conselhos sobre informações relacionadas à saúde [2, 3]. Essas fontes contêm crenças gerais de saúde pública e têm o potencial de ampliar a compreensão de tópicos como diagnóstico, medicamentos e alegações. Existem quase 140 usos potenciais do Twitter na área da saúde [4], os usos mais comuns são: alerta e resposta a desastres, gerenciamento de diabetes, alertas de segurança de medicamentos da Food and Drug Administration, captura e relatórios de dados de dispositivos biomédicos, licitação de turnos para enfermeiros e outros profissionais de saúde, brainstorming de diagnóstico, rastreamento de doenças raras e conexão de recursos, assistência para parar de fumar, dicas de cuidados infantis para novos pais e consultas de pacientes pós-alta e cuidados de acompanhamento [4].

Como exemplo de como as Redes Sociais retratam as informações médicas, mesmo quando a segurança e eficácia no A vacina contra o papilomavírus humano (HPV) foi comprovada, as tendências da Rede Social relatam baixa eficácia em alguns países, incluindo os Estados Unidos. No entanto, essas opiniões e informações negativas são induzidas por notícias, celebridades ou criadores de tendências e impactam a confiança do público neste tópico específico [5].

Devido ao impacto das redes sociais na saúde há um interesse crescente no desenvolvimento de modelos e sua análise. Prieto et. al. (2014) apresentam a análise do valor dos tweets relacionados à saúde, este estudo utiliza técnicas de aprendizado de máquina para avaliar esses tweets; eles coletam os dados com base na expressão regular na Espanha e em Portugal, então eles restringem o documento a quatro categorias selecionadas “Gravidez”, “Depressão”, “Gripe” e “Transtorno Alimentar” eles utilizam dois métodos tradicionais de aprendizado de máquina KNN, SVM [6].

Prier et. al. (2011) propõem um modelo baseado no modelo LDA e definem o modelo para gerar 250 tópicos, selecionam “Tabaco” como tópico para validar o modelo [7]. Dois outros estudos, um no Reino Unido e outro nos EUA, encontraram a correlação entre a análise de sentimento do twitter e a qualidade dos serviços de saúde [8,9].

Neste trabalho, é apresentado um método automatizado de modelagem de tópicos do Twitter. Este sistema não será semeado ou inicializado e melhorará a partir de feedback positivo e negativo. O sistema desenvolvido coleta os tweets e utiliza a Alocação de Dirichlet Latente (LDA) como modelo não supervisionado, rotulando cada tweet, identificando padrões. Este método destina-se a processar tweets relacionados a crenças públicas sobre saúde. Projetamos uma CNN combinada com o modelo Word2Vect. O modelo Word2Vect foi treinado em 7.821 resumos médicos como uma primeira iteração de aprendizado. Os resultados deste treinamento enriquecem o vocabulário relacionado à saúde, aprimoram o método de detecção de tweets relacionados e aprimoram a modelagem de tópicos gerais para detecção de novos tweets.

II. METODOLOGIA

Os dados de texto na área da saúde podem ser categorizados em três domínios Clínico, Biomédico e Social, cada um deles reunido por um grupo separado de pessoas. O texto biomédico é coletado por cientistas e médicos com experiência em laboratório. As notas clínicas geradas pelo pessoal médico referem-se a um paciente específico, o texto biomédico, por outro lado, refere-se à população geral de pacientes.
O texto da rede social fornecido está relacionado com uma ideia, conselhos de pessoas ou informações específicas, mas a veracidade dessas informações não é garantida.

Neste estudo, os dados de tweets relacionados à saúde foram coletados por um mês, e a correlação de tópicos de hashtag foi modelada usando a técnica LDA. Também implementamos um método para

978-1-5386-7568-7/18/$31,00 ©2018 IEEE

detectar novos documentos relacionados com os temas para recolher dados futuros.

A. Conjunto de dados

Foi coletado um conjunto de dados de 144.922 tweets em inglês, as palavras-chave empregadas foram: saúde, saúde, médicos, atendimento domiciliar, saúde digital e saúde digital. Durante a coleta de dados foram coletados diversos tweets relacionados a ofertas de emprego, pois esses dados estão fora do escopo do presente estudo, foram excluídos os tweets contendo as palavras-chave: job, Job, (hir\w+), career e admission. Neste trabalho nosso interesse era coletar informações dentro dos EUA, portanto outra limitação foi imposta, os tweets foram filtrados com base no nome do estado dos EUA, nome do condado e Publicação Padrão de Processamento de Informação Federal conhecida como código FIPS.

Depois de filtrar os dados, 37.910 tweets relacionados à saúde atendem aos critérios. O conjunto de dados final contém tweets de 43 Estados dos EUA, sendo a Califórnia com 5.923 tweets o mais populoso e Dakota do Sul com 43 tweets o menos populoso. Os dados foram coletados a partir de outubro de 2018 durante um mês.

B. Preparação de Dados

Para preparar os dados para o processamento, o primeiro passo é remover caracteres de linha, como caracteres de entrada e tabulações, e remover todas as aspas, hashtags, números e não caracteres usando padrões de expressão regular. Além disso, todos os URLs ("https?://[A-Za-z0-9./]+") e referências (@[A-Za z0-9]+) foram removidos com padrões de expressão regular.

A segunda etapa consiste na conversão de todo o texto para tokens, para isso foi empregada a biblioteca Genism [10].
Em seguida, as palavras de parada como "a, an, the" foram removidas, essas palavras não agregam nenhum valor à análise do texto, adicionando ruído aos dados.

A última etapa é o processo de Lematização e Stemming, que permite obter as características necessárias. Usando Stemming, as terminações flexionais, prefixo e sufixo das palavras são removidos. Com a Llematização, é realizada uma análise morfológica das palavras, este método requer a predefinição de um vocabulário para a língua alvo; esse processo foi conduzido com base em um dicionário fornecido pelo genism [10]. Neste sistema as palavras selecionadas foram 'NOUN', 'ADV', 'ADJ' e 'VERB'.

O pré-processamento dos dados foi realizado por meio de duas listas: a primeira lista denominada "V" é o vocabulário do texto contendo cada palavra com sua frequência correspondente, e a segunda denominada "W" é a lista das palavras tokenizadas para cada documento. Com base em W e V foi criado um Bag of Words (BOW). Portanto, os documentos são representados por uma lista de vetores de comprimento "V". A partir do pré-processamento obtém-se uma matriz com cada documento em suas linhas e o vocabulário nas colunas.

C. Detecção Automatizada de Tópicos

A modelagem de tópicos consiste em encontrar padrões ou palavras relevantes dentro de um pacote de documentos não rotulados. Para realizar esta tarefa foi implementado um LDA baseado em um modelo Bayesiano hierárquico de três níveis [11].

O LDA foi treinado em três bancos de dados não rotulados. Este modelo permite encontrar onde os dados são mais densos, assim definir o tópico. Assim como nas técnicas não supervisionadas, o desafio é definir o número de clusters que irão representar os tópicos; isso implica definir uma métrica para a densidade, levando ao número ótimo de tópicos (clusters) no corpus. Duas métricas do Natural Language Processing foram empregadas, Perplexity e log-likelihood definidas em (1) e (2) respectivamente.

ÿ ÿ ÿ (1)

Onde M é o número de documentos, p( ) é a probabilidade de uma dada palavra wd, e Nd é o número total de palavras por documento.

O (2)

Onde P é a probabilidade condicional de uma dada palavra wd.

Os valores extremos dessas métricas são selecionados como o melhor número de tópicos para o corpus. O modelo LDA foi aplicado para diferentes números de componentes, este número é interpretado como o melhor número de tópicos que descrevem o corpus. O modelo foi executado para 2, 5, 10, 20, 50, 100, 200 tópicos. Dadas as métricas, cinco foi o melhor número de tópicos com uma Perplexidade de 798,806 e valor de -442786,843 para o log-Probabilidade.

Para validar essas métricas foi realizada a Análise de Componentes Principais (ACP). Na Figura 1 é apresentado o mapa de distância da modelagem do tópico obtido pelo PCA. Com este modelo pode-se observar que utilizando os cinco tópicos selecionados, sem sobreposição; nesta figura o tamanho de cada círculo representa a população de cada tópico do corpus. A Tabela I apresenta a distribuição de cada tópico.

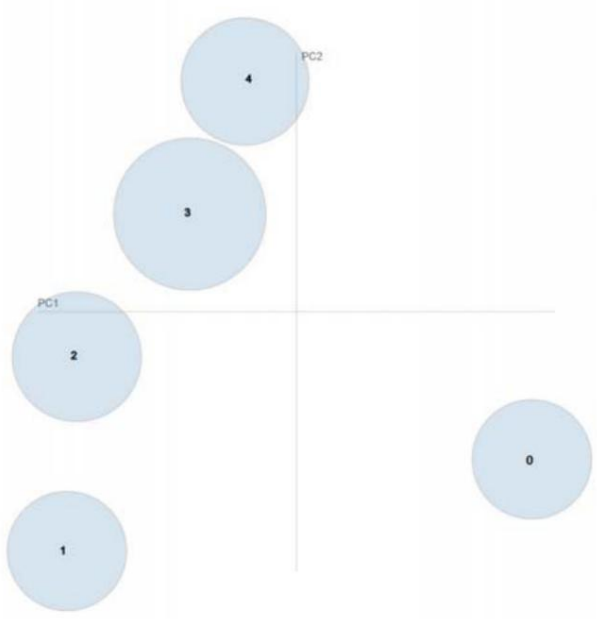

Figura 1 Mapa de distância entre tópicos

Representamos os termos mais relevantes com base no "LDAvis" uma ferramenta baseada na web [17]. Nas Figuras 2 a 6 são apresentados os 30 termos mais relevantes dentro de cada tópico e respectivas frequências, as barras azuis representam a frequência daquele termo em todos os documentos e as barras vermelhas são ordenadas pela relevância dos termos dentro de cada tópico.

TABELA I. FREQUÊNCIA E PROPORÇÃO DE CADA TÓPICO EM 37.910 TWEETS COLETADOS

Tópico	Frequência	Proporção
0	7.459	19,68
1	6.070	16.01
2	7.994	21.09
3	9.919	26.16
4	6.468	17.06
TOTAL	37.910	

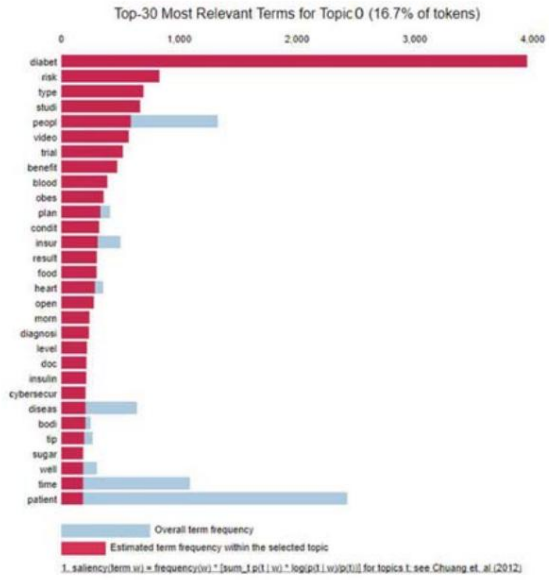

Figura 2 Os 30 principais termos mais relevantes para o tópico 0

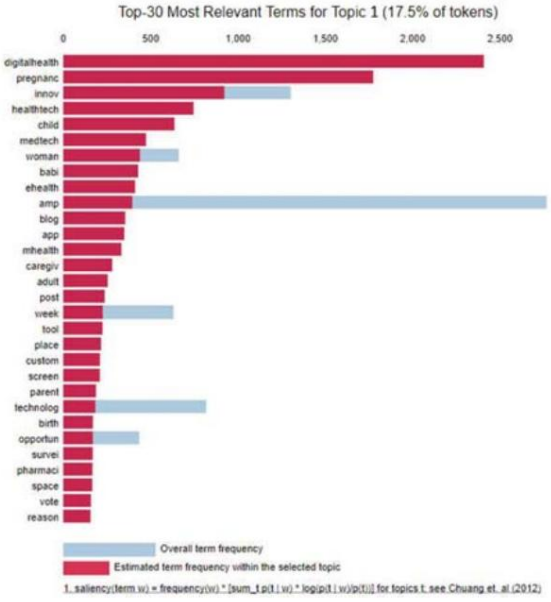

Figura 3 Os 30 principais termos mais relevantes para o tópico 1

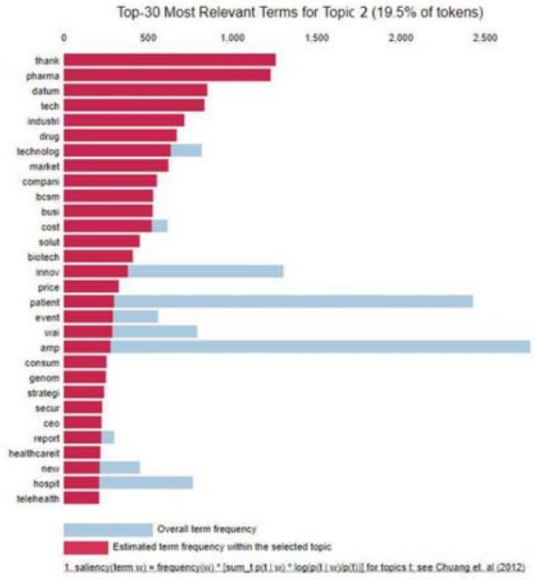

Figura 4 Os 30 principais termos mais relevantes para o tópico 2

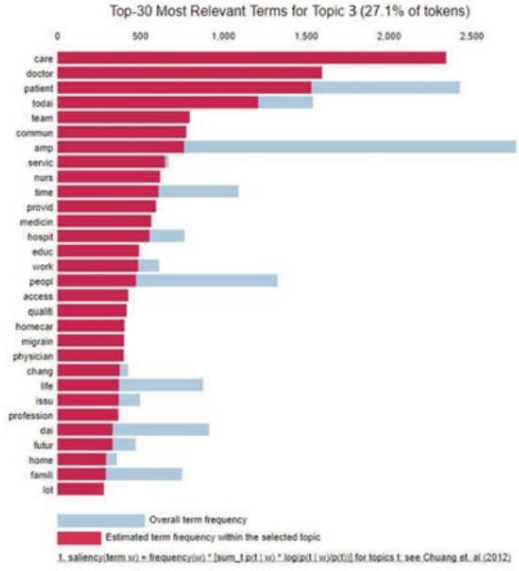

Figura 5 Os 30 principais termos mais relevantes para o tópico 3

D. Métricas de previsão

LDA usado como modelo não supervisionado para categorizar os tweets. Cinco é o número ideal de tópicos com base nas métricas de perplexidade e probabilidade de log. Os tópicos selecionados descritos na Tabela II são diabéticos, saúde digital, mercado de medicamentos, serviços de saúde e câncer e pesquisa. LDA atribui 5 pontuações (número de tópicos) a um tweet, cada uma delas representa a pontuação de similaridade com tópicos pré-definidos, então selecionamos a pontuação mais alta e a rotulamos como o Tópico correspondente. Para validar nossos resultados, usamos rotulagem LDA como verdade e comparamos com rótulos previstos. Portanto, utilizar o modelo LDA transfere os dados de uma técnica não supervisionada para uma técnica supervisionada.

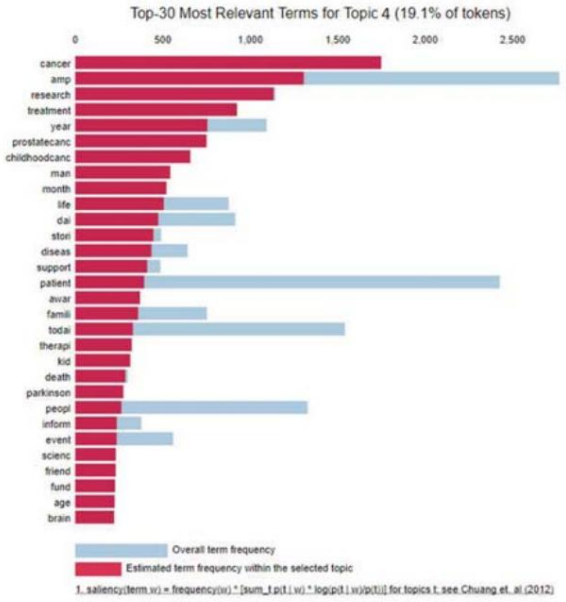

Figura 6 Os 30 principais termos mais relevantes para o tópico 4

TABELA II. TÓPICO GERADO COM BASE NA LDA MODELO

Tópico 0	Tópico 1	Tópico 2	Tópico 3	Tópico 4
diabetes	digitalsaúde	agradecer	Cuidado	Câncer
risco	gravidez	farmacêutico	doutor	amplificador
tipo	inovação	dado	paciente	pesquisa
estudar	tecnologia da saúde	tecnologia	hoje	tratamento
pessoas	criança	indústria	equipe	ano
vídeo	tecnologia médica	medicamento	câncer de próstata	comum
tentativas	mulher	tecnologia	amplificador	infânciacanc
beneficiar	babi	mercado	serviço	homem
sangue	esaúde	companhia	enfermeiras	mês
obesos	amplificador	bcsm	Tempo	vida

Os resultados da classificação são avaliados por quatro métricas: exatidão (3), recall (4), precisão (5) e F1-score (6), nestas equações TP é verdadeiro positivo, TN é verdadeiro negativo, FP é falso positivo e FN é falso negativo. A precisão representa a exatidão do classificador e mede quantos rótulos previstos estão relacionados, o recall representa quantos registros verdadeiros são previstos e os F-scores quantificam a média harmônica da precisão e do recall.

ÿ ÿÿ ÿ ÿÿ / ÿÿÿ ÿ ÿ ÿ ÿÿ ÿ ÿÿÿÿ (3)

ÿÿ ÿÿ / ÿÿ ÿ ÿÿÿ (4)

ÿÿÿÿÿÿ ÿÿ / ÿÿÿ ÿ ÿÿÿ ÿ (5)

ÿ ÿ ÿ *f f* ÿÿÿÿ (6 (e)

E. Arquitetura e resultados de classificação O modelo proposto é construído por camadas. Na primeira camada os dados são coletados usando uma ferramenta python chamada Teewpy [12]. A segunda camada contém os métodos de limpeza e pré-processamento descritos, convertendo os tweets em vetores que podem ser processados. A terceira camada é um método Word2Vec que cria uma matriz com base no vetor recebido pela última camada e usa essa matriz para inicializar uma rede neural para prever o tweet rotulado.

Um classificador CNN é a quarta camada, onde os tweets não vistos vindos do Word2Vec são rotulados. Normalmente, uma modelagem de sequência está relacionada à Rede Neural Recorrente (RNN), porém os resultados indicam uma perspectiva diferente. A Rede Neural Convolucional (CNN) fornece resultados notáveis em PNL [13]. Yih et ai. 2011 aplicou a CNN na análise semântica [14], Shen 2014 a utilizou para recuperação de consultas [15], KalchBanner 2014 a usou para modelagem de frases [16] e Yoon Kim 2014 conectou o modelo Word2Vec com a CNN [13]. Como parte deste projeto de pesquisa, estávamos explorando modelos de classificação que podemos usar como parte de um sistema adaptativo. A seleção de recursos é um dos desafios e as Redes Neurais Convolucionais (CNN) são apropriadas, pois não exigem seleção de recursos a priori. Uma limitação das CNNs é que elas exigem um tamanho de entrada fixo, pois os tweets são limitados a 280 caracteres, podemos usar preenchimento para tweets mais curtos preservando os tamanhos de entrada fixos. Nossa arquitetura compreende três camadas convolucionais com tamanhos de kernel de 128, 64 e 32, respectivamente. O sistema tem uma queda de 0,5. Ele foi iterado por 200 épocas usando lotes de 100 registros com uma taxa de aprendizado de 0,00001. Os pesos foram obtidos usando Adam-Optimizer.

Nesta camada alguns tweets podem não se encaixar nos tópicos selecionados, estes são sinalizados como dados não rotulados e armazenados em um conjunto de dados independente. Se os tweets se encaixam nos tópicos selecionados, eles são classificados e rotulados de acordo.

Na quinta camada, dados não rotulados são alimentados para um modelo LDA e novos tópicos podem ser criados. O modelo CNN será treinado novamente, atualizando tanto os tópicos quanto o modelo Word2Vec.

Na Figura 7 é apresentada a arquitetura do sistema. Essa arquitetura visa proporcionar um aprendizado ativo de tópicos e reforçar a classificação do modelo CNN.

Para comparar o sistema proposto, as técnicas SVM e CNN foram implementadas de forma independente para prever e rotular novos tweets. No entanto, as capacidades de previsão de ambos os métodos foram limitadas devido aos conjuntos de dados desequilibrados. Os resultados utilizando as métricas de previsão são apresentados na Tabela II.

TABELA III. COMPARAÇÃO DE MODELOS

Precisão do algoritmo	SVM 39,5%	CNN 57%		Pontuação F
CNN-estático	83,34%	67,6%		34,9%
		58,8%		56%
		83%	Relembrar 39,5%	83,34% 84%

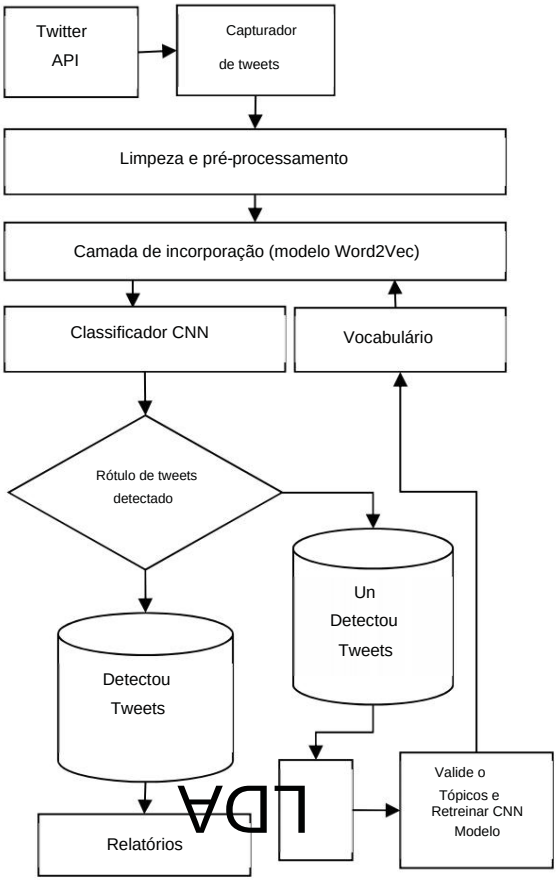

Figura 7 Arquitetura proposta para rastrear os tópicos do Twitter

III. RESULTADOS

Os tweets coletados foram processados usando o modelo LDA, e com as métricas de perplexidade e log-verossimilhança os tópicos foram definidos, a Tabela III apresenta as 10 palavras-chave de cada um dos tópicos. Um exemplo da classificação dos tweets coletados é apresentado na Tabela IV, onde é apresentada a porcentagem de similaridade para cada tópico.

Após a modelagem de tópicos, representamos a distribuição de cada tópico nos diferentes estados usando um mapa de calor. As áreas vermelhas representam os estados onde mais pessoas estão interessadas nesse tópico. Este relatório oferece a possibilidade de retratar a evolução temporal de um tema nos diferentes Estados.

Cada tweet pertence a um local, representado pela abreviatura correspondente do nome do Estado. Para calcular a porcentagem de interesse de cada estado em um tópico específico, dividimos a população de tweets naquele estado pela população do tópico específico naquele estado.

Na Figura. 8 pode-se observar que a Califórnia é o estado mais interessado no tópico 0 “Diabéticos” com 85,22% e Virgínia com 30,98% é o menos interessado. Figura. 9 mostra que Nova York é o estado mais interessado no tópico 1 “Saúde Digital” com 77,3% e Wisconsin com 29,13% o menos interessado.

No caso do tópico 2, a palavra mais frequente não pode ser utilizada como rótulo do tópico, mas as palavras contidas neste tópico permitem rotulá-lo como “Mercado Medicamento e Farmácia”. Em Figura. 10 é apresentada a distribuição para este tópico, aqui a Califórnia é o estado mais interessado com 100% e Oklahoma com 21,89% o menos interessado.

TABELA IV. EXEMPLO DE TWEET E TÓPICO SEMELHANÇA BASEADA NO MODELO LDA

Exemplos de Tweets	0	1	2	3	4
As mulheres grávidas devem Coma Mais Peixe	10%	59%	10%	10%	10%
Saúde examina Ho w para proteger contra novos Mobilidade e #IoT Secur Ameaças na #Saúde por Tech #cibersegurança # healthdata #healthIT #HI T	55%	5%	30%	5%	5%
Reenquadramento e endereçamento g violência horizontal como melhoria da qualidade do local de trabalho preocupação com o movimento #digitalh saúde #inovação #saúde tecnologia #medtech #li #fb	3%	22%	3%	50%	20%
Um paciente com IMC de 32 com #doençametabólica #diabetes #hipertensão alta #colesterol etc wou Preciso de #CirurgiaBariátrica amp #MetabolicSurgery mais do que um paciente com #IMC 37 sem esses c condições Infelizmente um n limite de IMC desatualizado de 35 ainda é nosso critério'# healthtech	40%	10%	2%	15%	33%
como a Casa Branca Tentativas de projeto do lthcare para injetar valor no U S #sistema de saúde #pb m #preçosdedrogas #farma'	20%	22%	51%	3%	3%

O tópico 3 “Cuidados” é mostrado na Figura. 11. Califórnia e Nova York são as mais interessadas neste tema com 80% e Virgínia com 41% é a menos interessada. O último tópico representado pelos rótulos “câncer”, “amputação” e “pesquisa” representados na Figura. 12, mostra que há um interesse generalizado sobre esses temas, sendo Califórnia, Texas, Nova York, Wisconsin e Ohio os estados mais interessados.

4. DISCUSSÃO E CONCLUSÕES

Neste artigo é apresentada a implementação de um sistema de análise de tweets. O sistema compreende desde o processo de coleta até a classificação dos novos documentos coletados. Para classificar os tweets foi implementada uma Rede Neural Convolucional (CNN) em conjunto com um modelo Word2Vect.
Este sistema fornece feedback aos tópicos, permitindo a geração de novos tópicos. O algoritmo de classificação tem acurácia de 83,34%, precisão de 83%, recall de 84% e F Score de 83,8%. As tendências sobre os tópicos analisados com este sistema podem ser retratadas e relatadas de forma localizada, neste artigo foram usados tweets dos EUA como Estudo de Caso, os mapas de calor correspondentes a essas tendências foram apresentados e descritos.

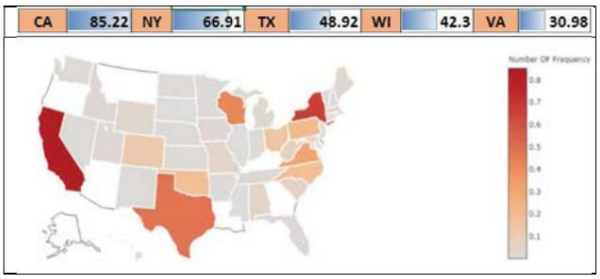

Figura 8 Distribuição do Tópico 0 nos Estados Unidos

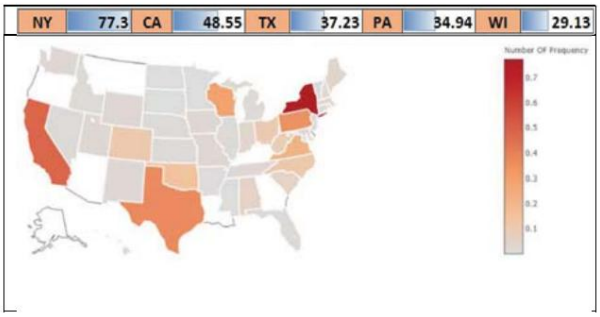

Figura 9 Distribuição do Tópico 1 nos Estados Unidos

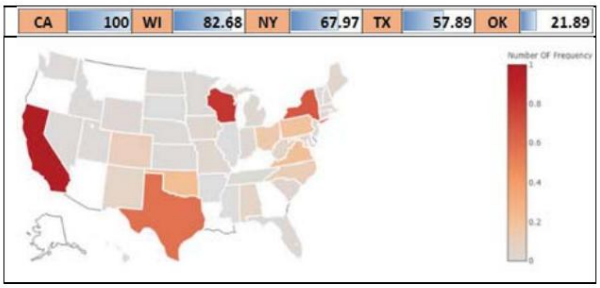

Figura 10 Distribuição do Tópico 2 nos Estados Unidos

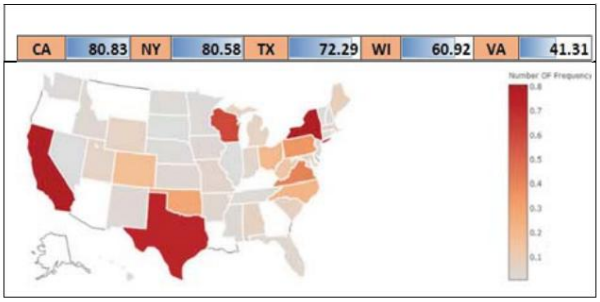

Figura 11 Distribuição do Tópico 3 nos Estados Unidos

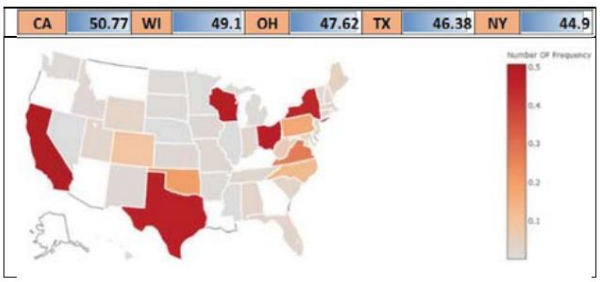

Figura 12 Distribuição do Tópico 4 nos Estados Unidos

REFERÊNCIAS

[1] Reddy, CK, & Aggarwal, CC (2015). Análise de dados de saúde. Chapman e Hall/CRC.

[2] Scanfeld, D., Scanfeld, V., & Larson, EL (2010). Divulgação de informações de saúde através das redes sociais: Twitter e antibióticos. Jornal americano de controle de infecção, 38(3), 182-188.

[3] Prier, KW, Smith, MS, Giraud-Carrier, C., & Hanson, CL (2011, março). Identificando tópicos relacionados à saúde no twitter. Em Conferência internacional sobre computação social, modelagem comportamental-cultural e previsão (pp. 18-25). Springer, Berlim, Heidelberg.

[4] Chapman, BE, Lee, S., Kang, HP e Chapman, WW (2011). Classificação em nível de documento de relatórios de angiografia pulmonar por TC com base em uma extensão do algoritmo ConText. Jornal de informática biomédica, 44(5), 728-737.

[5] Surian, D., Nguyen, DQ, Kennedy, G., Johnson, M., Coiera, E., & Dunn, AG (2016). Caracterizando discussões no Twitter sobre vacinas contra o HPV usando modelagem de tópicos e detecção da comunidade. Journal of medical Internet research, 18(8).

[6] Prieto, VM, Matos, S., Alvarez, M., Cacheda, F., & Oliveira, JL (2014). Twitter: um bom lugar para detectar condições de saúde. PloS um, 9(1), e86191.

[7] Prier, KW, Smith, MS, Giraud-Carrier, C., & Hanson, CL (2011, março). Identificando tópicos relacionados à saúde no twitter. Em Conferência internacional sobre computação social, modelagem comportamental-cultural e previsão (pp. 18-25). Springer, Berlim, Heidelberg.

[8] Greaves, F., Laverty, AA, Cano, DR, Moilanen, K., Pulman, S., Darzi, A., & Millett, C. Tweets sobre qualidade hospitalar: um estudo de métodos mistos. BMJ Qual Saf. 2014 outubro; 23 (10): 838–46. doi: 10.1136/bmjqs 2014-002875.

[9] Hawkins, JB, Brownstein, JS, Tuli, G., Runels, T., Broecker, K., Nsoesie, EO, ... & Greaves, F. (2015). Medindo a qualidade de atendimento percebida pelo paciente em hospitais dos EUA usando o Twitter. BMJ Qual Saf, bmjqs 2015.

[10] Rehurek, R., & Sojka, P. (2010). Framework de software para modelagem de tópicos com grandes corpora. Em Anais do Workshop LREC 2010 sobre Novos Desafios para Estruturas de PNL.

[11] Blei, DM, Ng, AY, & Jordan, MI (2003). Alocação de dirichlet latente. Journal of Machine Learning Research, 3 (Jan), 993-1022.

[12] Roesslein, J. (2015). Tweepy. Módulo de linguagem de programação Python.

[13] Kim, Y. (2014). Redes neurais convolucionais para sentença classificação. arXiv pré-impressão arXiv:1408.5882.

[14] Yih, WT, Toutanova, K., Platt, JC, & Meek, C. (2011, junho). Aprendizagem de projeções discriminativas para medidas de similaridade de texto. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning (pp. 247-256). Associação de Linguística Computacional.

[15] Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014, abril). Aprendendo representações semânticas usando redes neurais convolucionais para pesquisa na web. In Proceedings of the 23rd International Conference on World Wide Web (pp. 373-374). ACM.

[16] Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). Uma rede neural convolucional para modelagem de sentenças. arXiv pré-impressão arXiv:1404.2188.

[17] Sievert, C., & Shirley, K. (2014). LDAvis: Um método para visualizar e interpretar tópicos. Em Anais do workshop sobre aprendizagem interativa de línguas, visualização e interfaces (pp. 63-70).