Tendências em Saúde nas Mídias Sociais: Análise usando o Twitter Modelagem de Tópicos

Mohsen Asghari
Departamento de Engenharia Informática
e Ciências da Computação
CECS Universidade de Louisville Louisville, KY, EUA
m0asgh02@louisville.edu

Daniel Sierra-Sosa
Departamento de Engenharia Informática
e Ciências da Computação
CECS Universidade de Louisville Louisville, KY, EUA
d.sierrasosa@louisville.edu

Adel Elmaghraby
Departamento de Engenharia Informática
e Ciências da Computação
CECS Universidade de 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. Estes posts contêm opiniões públicas sobre saúde, e permitem entender como é a percepção popular sobre tópicos como diagnóstico médico, medicamentos, instalações e reivindicações. Neste artigo, apresentamos um sistema adaptativo projetado utilizando 5 camadas. O sistema contém uma combinação de algoritmos não supervisionados e supervisionados para acompanhar as tendências das mídias sociais de saúde. Por ser baseado em um modelo word2vec, ele também captura a correlação de palavras com base no contexto, melhorando ao longo do tempo, aumentando a precisão das previsões e o rastreamento de tweets. Neste trabalho, nos concentramos nos dados dos Estados Unidos e os usamos para detectar os tópicos de tendências de cada estado. Estes tópicos são seguidos, incluindo novas contribuições de redes sociais. O algoritmo supervisionado implementado é um Convolutional Neural Network (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 uma precisão 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 que demonstram uma vantagem que pode ser aproveitada para melhorias adicionais.

Palavras-chave - Tweets de Saúde, LDA, Classificação, Aprendizagem Profunda

I. INTRODUÇÃO

O texto livre em saúde é classificado em dois grupos: texto Biomédico e texto Clínico. Os textos biomédicos incluem livros, resumos e artigos. O texto clínico inclui relatórios de pessoal médico, como as patologias diagnosticadas do paciente, histórico pessoal e médico [1]. Entretanto, textos relacionados à saúde também estão disponíveis na rede social como texto livre/escondicionado, constituindo ainda mais um grupo de textos relacionados à saúde.

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 expandir a compreensão de tópicos como diagnóstico, medicamentos e reivindicações. Há quase 140 usos potenciais da saúde no Twitter [4], os usos mais comuns são: alerta e resposta a desastres, gerenciamento de diabetes, alertas de segurança de medicamentos da Food and Drug Administration, coleta de dados e relatórios de dispositivos biomédicos, licitação de turno para enfermeiros e outros profissionais da saúde, brainstorming de diagnóstico, rastreamento de doenças raras e conexão de recursos, assistência à cessação do fumo, dicas de cuidados infantis para novos pais e consultas e acompanhamento de pacientes após a alta [4].

Como um exemplo de como as Redes Sociais retratam a medicina
informação, mesmo quando a segurança e a eficácia no

978-1-5386-7568-7/18/$31.00 ©2018 IEEE

A vacina contra o papilomavírus humano (HPV) tem sido comprovada, as tendências da Rede Social relatam baixa eficácia em alguns países, incluindo os Estados Unidos. No entanto, esta opinião e informação negativa é induzida por notícias, celebridades ou criadores de tendências, e elas impactam a confiança do público neste tópico em particular [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 a partir do valor dos tweets relacionados à saúde, este estudo utiliza técnicas de aprendizagem mecânica para avaliar esses tweets; eles coletam os dados com base na expressão regular na Espanha e em Portugal, depois restringem o documento a quatro categorias selecionadas "Gravidez", "Depressão", "Gripe" e "Desordem Alimentar" utilizaram dois métodos tradicionais de aprendizagem mecânica KNN, SVM [6].

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

Neste trabalho, é apresentado um método automatizado de modelagem de tópicos no Twitter. Este sistema não será semeado ou inicializado e melhorará a partir do feedback positivo e negativo. O sistema como desenvolvido coleta tweets e pelo uso da Latent Dirichlet Allocation (LDA) como modelo não supervisionado, etiquetando cada tweet, identificando padrões. Este método se destina a processar tweets relacionados às crenças públicas sobre saúde. Projetamos um modelo CNN combinado com o 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 com a saúde, melhoram o método de detecção de tweets relacionados e melhoram a modelagem geral do tópico para detecção de novos tweets.

II. METODOLOGIA

Os dados de texto em saúde podem ser categorizados em três domínios Clínico, Biomédico e Social, cada um deles reunido por um grupo separado de pessoas. Os textos biomédicos são coletados por cientistas e médicos que têm experiência em laboratório. As notas clínicas geradas pelo pessoal médico se referem a um paciente específico, o texto biomédico, por outro lado, se refere à população geral de pacientes. O texto da rede social fornece uma idéia, conselhos de pessoas ou informações específicas, mas a veracidade destas informações não é garantida.

Neste estudo foram coletados dados de tweet relacionados à saúde durante um mês, e a correlação temática hashtag foi modelada usando a técnica LDA. Também implementamos um método para

detectar novos documentos relacionados com os tópicos, a fim de coletar dados futuros.

A. Dataset

Foi coletado um conjunto de 144.922 tweets em inglês, as palavras-chave empregadas foram: healthcare, health, doctors, homecare, digitalhealth, e digital health. Durante a coleta de dados foram coletados vários tweets relacionados com ofertas de emprego, pois estes dados estão fora do escopo do estudo atual, foram excluídos tweets contendo as palavras-chave: job, Job, (hir\w+), carreira e contratação. 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 americano, nome do condado e Publicação Padrão Federal de Processamento de Informações conhecida como código FIPS.

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

B. Preparação de dados

A fim de preparar os dados para o processamento, o primeiro passo é remover os caracteres da linha, tais como digitar caracteres e tabulações, depois remover todas as aspas, hashtags, números e não caracteres, usando padrões de expressão regulares. Além disso, todas as URLs ("*https?://[A-Za-z0-9./]+*") e referências *(@[A-Za- z0-9]+*) foram removidas com padrões de expressão regulares.

O segundo passo consiste na conversão de todo o texto em fichas, para este fim foi utilizada a biblioteca do genismo [10]. Em seguida, as palavras de parada como "*a, an, the*" foram removidas, estas palavras não acrescentam nenhum valor à análise do texto, acrescentando ruído aos dados.

A última etapa é o processo de Lemmatização e Corte, que permite a obtenção das características necessárias. Ao utilizar o Stemming, as terminações inflecionais, prefixo e sufixo das palavras são removidos. Com a Llemmatização, é realizada uma análise morfológica das palavras, este método requer a predefinição de um vocabulário para a língua-alvo; este processo foi conduzido com base em um dicionário fornecido pelo genismo [10]. Neste sistema as palavras selecionadas foram 'NOUN', 'ADV', 'ADJ' e 'VERB'.

O pré-processamento de dados foi realizado utilizando duas listas: a primeira lista chamada "V" é o vocabulário do texto contendo cada palavra com sua freqüência correspondente, e a segunda chamada "W" é a lista das palavras simbólicas para cada documento. Com base em W e V, foi criado um Saco de Palavras (BOW). Portanto, os documentos são representados por uma lista de vetores com o comprimento de "V". A partir do pré-processamento é obtida uma matriz com cada documento em suas fileiras e o vocabulário nas colunas.

C. Detecção Tópica Automatizada

A modelagem do tópico consiste em encontrar padrões ou palavras relevantes dentro de um saco de documentos não etiquetados. Para realizar esta tarefa, foi implementado um LDA baseado em um modelo Bayesiano de três níveis hierárquicos [11].

A LDA foi treinada em três bancos de dados não rotulados. Este modelo permite encontrar onde os dados são mais densos, definindo assim o tópico. Como as técnicas não supervisionadas, o desafio é definir o número de clusters que representarão os tópicos;

isto implica definir uma métrica para a densidade, levando ao número ideal de tópicos (clusters) no corpus. Duas métricas do Processamento de Linguagem Natural foram empregadas, Perplexidade e Probabilidade de Log definido em (1) e (2) respectivamente.

(1) [illegible]

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

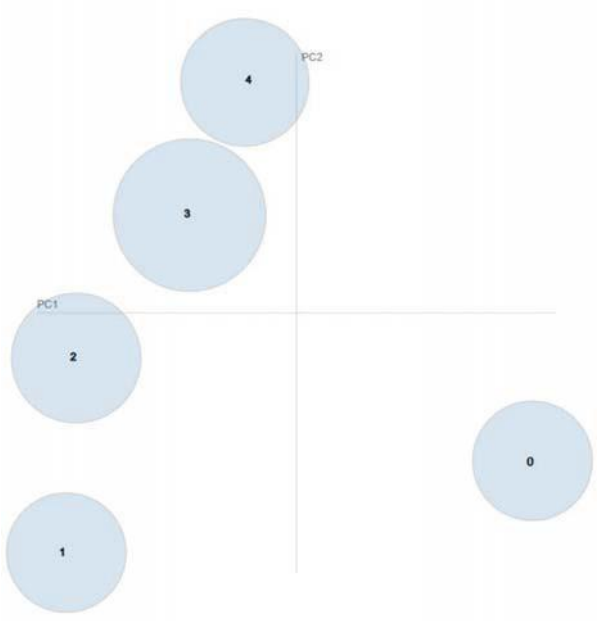
(2) [illegible]

W were *P* é a probabilidade condicional de uma determinada 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. Dada a métrica, cinco foi o melhor número de tópicos com uma Perplexidade de 798.806 e um valor de -442786.843 para a log-Likelihood.

Para validar estas métricas foi realizada a Análise de Componentes Principais (PCA). Na Figura 1 é apresentado o mapa de distância da modelagem do tópico obtido pela 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 no corpus. A tabela I apresenta a distribuição de cada tópico.

Figura 1 Mapa de Distância Inter-Tópica



Representamos os termos mais relevantes baseados em "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 freqüências, as barras azuis representam a freqüência desse termo sobre todos os documentos e as barras vermelhas são ordenadas pelos termos relevantes dentro de cada tópico.

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

Tema	Freqüência	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	

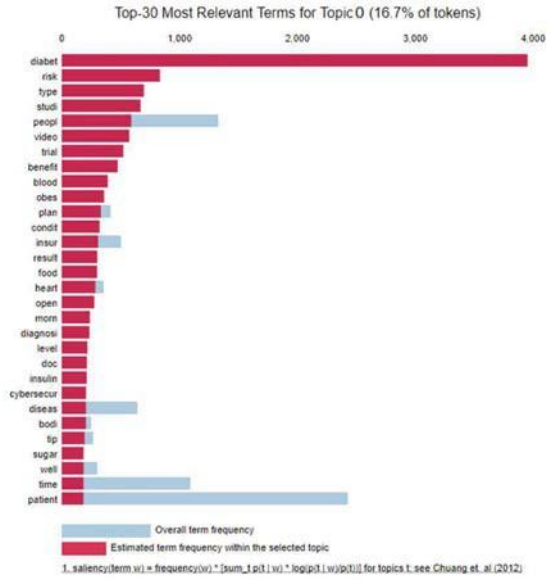


Figura 2 Top 30 Termos mais relevantes para o tópico 0

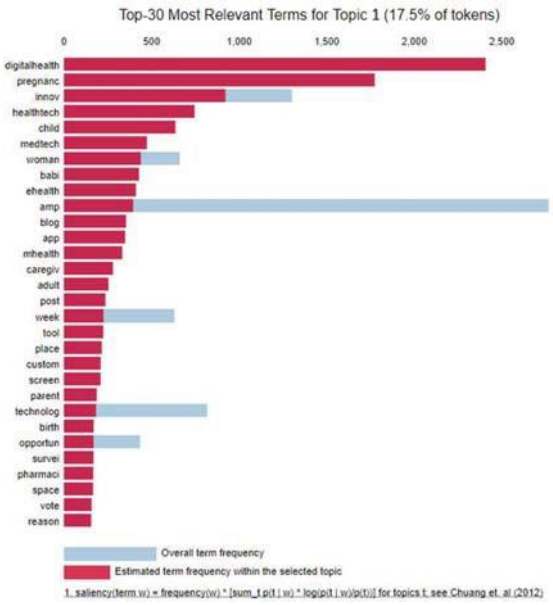


Figura 3 Top 30 Termos mais relevantes para o tópico 1

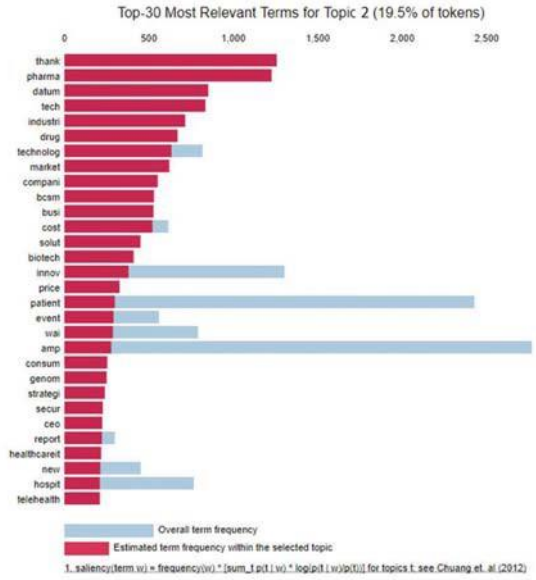


Figura 4 Top 30 Termos mais relevantes para o tópico 2

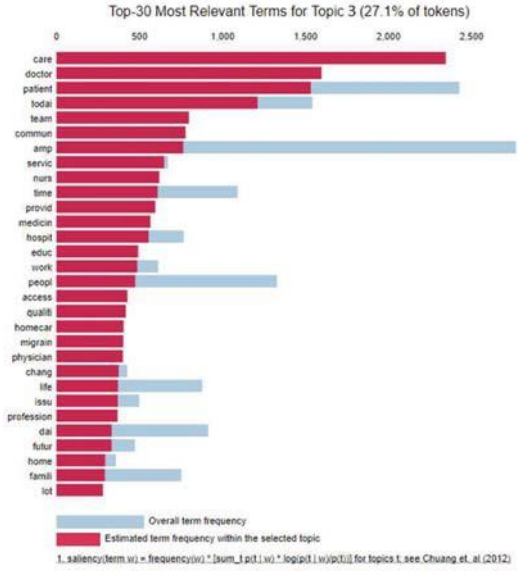


Figura 5 Top 30 Termos mais relevantes para o tópico 3

D. Métricas de Previsão

LDA usado como modelo não supervisionado para categorizar os tweets. Cinco é o número ideal de tópicos com base na perplexidade e nas métricas de probabilidade de log-likelihood. Os tópicos selecionados descritos na Tabela II são diabéticos, saúde digital, mercado de drogas, serviços de saúde e câncer e pesquisa. A LDA atribui 5 pontuações (número de tópicos) a um tweet, cada um deles representa a pontuação de similaridade com tópicos predefinidos, então selecionamos a pontuação mais alta e rotulamo-la como o Tópico correspondente. Para validar nossos resultados, usamos a etiquetagem LDA como verdade e a comparamos com as etiquetas previstas. Portanto, utilizando o modelo LDA, transferimos os dados de uma técnica não supervisionada para uma técnica supervisionada.

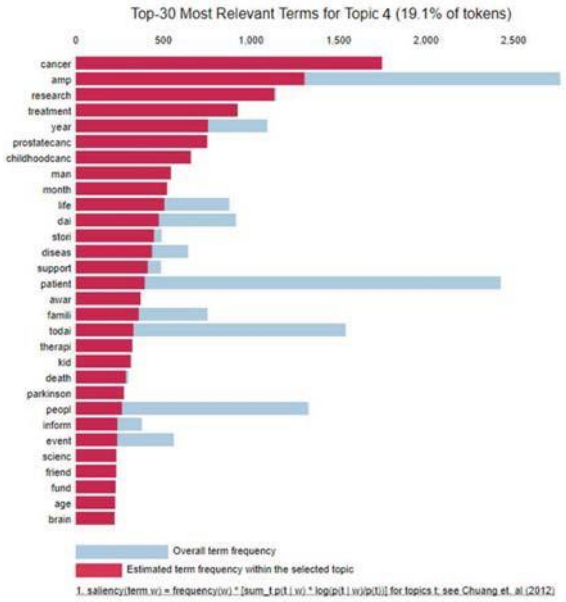


Figura 6 Top 30 Termos mais relevantes para o tópico 4

TABELA II.TÓPICO GERADO COM BASE NO MODELO LDA

Tópico 0	Tópico 1	Tópico 2	Tópico 3	Tópico 4
diabet	digitalhealth	obrigado	cuidados	câncer
risco	gravidez	farma	médico	amp
tipo	inovar	datum	paciente	pesquisa
studi	healthtech	tech	todai	tratamento
peopl	criança	industrial	equipe	ano
vídeo	medtech	droga	commun	prostatecanc
julgamento	mulher	tecnolog	amp	infânciacanc
benefício	babi	mercado	servic	homem
sangue	ehealth	compani	viveiro	mês
obes	amp	bcsm	tempo	vida

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

(3)

(4)

(5)

(6)

E. Arquitetura e resultados da 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 baseada no vetor recebido pela última camada e usa essa matriz para inicializar uma rede neural para prever o tweet etiquetado.

Um classificador CNN é a quarta camada, onde os tweets invisíveis provenientes do Word2Vec são etiquetados. Normalmente, uma modelagem de seqüência 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 al. 2011 aplicaram a CNN em análise semântica [14], Shen 2014 a utilizou para consulta de recuperação [15], KalchBanner 2014 a utilizou 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 utilizar como parte de um sistema adaptativo. A seleção de características é um dos desafios e as Redes Neurais Convolucionais (CNN) são apropriadas, pois não exigem uma seleção de características a priori. Uma limitação das CNNs é que elas requerem um tamanho de entrada fixo, já que os tweets são limitados a 280 caracteres que podemos usar para tweets mais curtos preservando os tamanhos de entrada fixos. Nossa arquitetura compreende três camadas convolutivas com tamanhos de 128, 64 e 32 kernel, respectivamente. O sistema tem uma queda de 0,5. Foi iterado por 200 épocas utilizando lotes de 100 registros com uma taxa de aprendizagem de 0,00001. Os pesos foram obtidos usando o Adam-Optimizer.

Nesta camada, alguns tweets podem não caber no selecionado
tópicos, estes são assinalados como dados não rotulados e armazenados em um conjunto de dados independente. Se os tweets se encaixarem nos tópicos selecionados, eles são classificados e rotulados de acordo.

Na quinta camada, os 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. Esta arquitetura visa proporcionar um aprendizado ativo dos tópicos e reforçar a classificação do modelo CNN.

Para comparar o sistema proposto, as técnicas SVM e CNN foram implementadas independentemente para prever e rotular novos tweets. Entretanto, as capacidades de previsão de ambos os métodos foram limitadas devido ao desequilíbrio dos conjuntos de dados. Os resultados usando a métrica de previsão são apresentados na Tabela II.

TABELA III. COMPARAÇÃO DE MODELOS

Algoritmo	Precisão	Precisão	Relembrar	F-Score
SVM	39.5%	67.6%	39.5%	34.9%
CNN	57%	58.8%	55.1%	56%
CNN-static	83.34%	83%	84%	83.8%

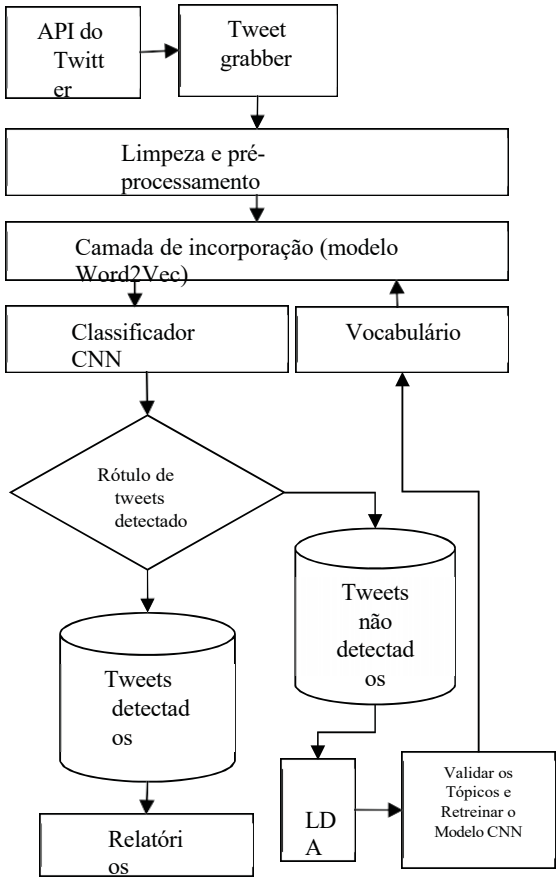


Figura 7 Arquitetura proposta para acompanhar os tópicos do Twitter

III. RESULTADOS

Os tweets coletados foram processados utilizando o modelo LDA,

e com a perplexidade e as métricas de probabilidade de logs, os tópicos
foram definidas, 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 a porcentagem de similaridade a cada tópico é retratada.

Após a modelagem dos tópicos, representamos a distribuição de cada tópico pelos diferentes Estados, utilizando um mapa térmico. As áreas vermelhas representam aqueles Estados onde mais pessoas estão interessadas nesse tópico. Este relatório oferece a possibilidade de representar a evolução temporal de um tópico sobre os diferentes Estados.

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

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

No caso do tópico 2, a palavra mais frequente não pode ser usada como o rótulo do tópico, mas as palavras contidas

Figura. 10 a distribuição para este tópico está representada, aqui a Califórnia é o estado mais interessado com um 100% e Oklahoma com 21,89% o menos interessado.

TABELA IV.EXEMPLO DE TWEET E SIMILARIDADE DE TÓPICOS COM BASE NO MODELO LDA

Amostra de Tweets	*0*	*1*	*2*	*3*	*4*
Mulheres grávidas devem comer mais peixe	10%	59%	10%	10%	10%
Saúde examina Ho w to Protect Against New Mobility e #IoT Secur i t y Threats in #Healthcare by Tech #cybersecurity # healthdata # healthdata #healthIT #HI T		5%	30%	5%	5%
Reframação e tratamento da violência horizontal como uma preocupação de melhoria da qualidade no local de trabalho #saúde digital #inovação #saúde tech #medtech #li #fb	3%	22%	3%	50%	20%
Um paciente com IMC de 32 com #doença metabólica #diabetes #hipertensão alta #colesterol, etc., necessitaria de #bariátrico-cirurgia mais do que um paciente com #IMC 37 sem estas condições Infelizmente, um limiar de IMC ultrapassado de 35 ainda é nosso critério'# healthtech	40%	10%	2%	15%	33%
como o plano da Casa Branca hea lthcare tenta injetar valor no sistema de saúde dos EUA #pb #forbes #farmácia".	20%	22%	51%	3%	3%

O tópico 3 "Cuidados" é mostrado na figura. 11. Califórnia e Nova Iorque são as mais interessadas neste tópico com 80% e

neste tópico permitem rotulá-lo como "Droga de Mercado e Farmácia". Em

Virgínia com 41% é a menos interessada. O último tópico representado pelos rótulos "câncer", "amputação" e "pesquisa" descritos na figura. 12, mostra que há um interesse generalizado nestes tópicos sendo Califórnia, Texas, Nova Iorque, Wisconsin e Ohio os estados mais interessados.

IV. Discussão e Conclusões

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

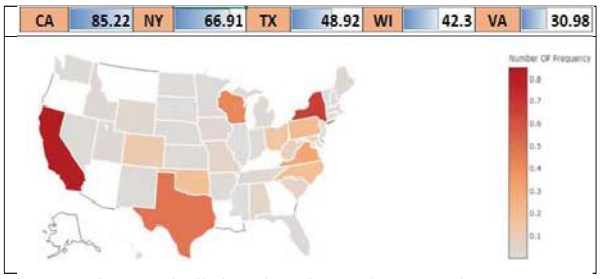
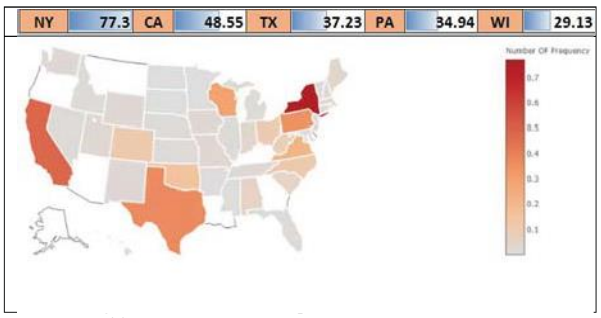
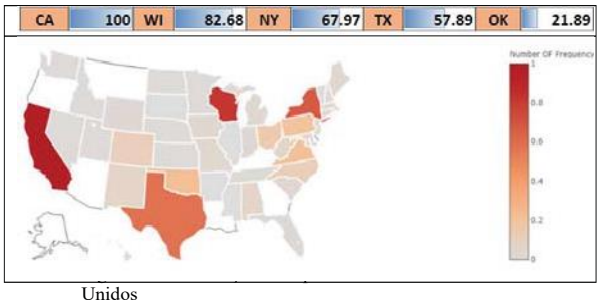


Figura 8 Distribuição do tópico 0 sobre os Estados Unidos



Unidos



Unidos

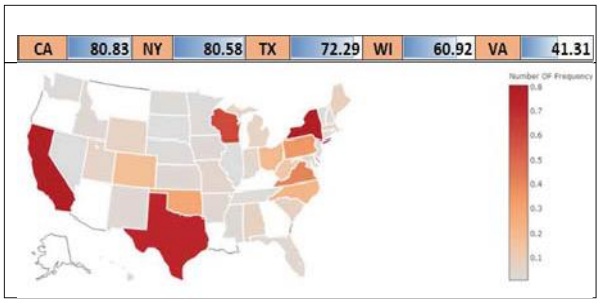


Figura 12 Distribuição do tópico 4 sobre os Estados Unidos

REFERÊNCIAS

[1] Reddy, C. K., & Aggarwal, C. C. (2015). Análise de dados da área de saúde. Chapman e Hall/CRC.

[2] Scanfeld, D., Scanfeld, V., & Larson, E. L. (2010). Disseminação de informações sobre saúde através de redes sociais: Twitter e antibióticos. Revista americana de controle de infecções, 38(3), 182-188.

[3] Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011, março). Identificação de tópicos relacionados à saúde no twitter. Em International conference on social computing, behavior-cultural modeling, and prediction (pp. 18-25). Springer, Berlim, Heidelberg.

[4] Chapman, B. E., Lee, S., Kang, H. P., & Chapman, W. W. (2011). Classificação em nível de documento dos relatórios de angiografia pulmonar por TC com base em uma extensão do algoritmo ConText. Journal of biomedical informatics, 44(5), 728-737.

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

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

[7] Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011, março). Identificação de tópicos relacionados à saúde no twitter. Em International conference on social computing, behavior-cultural modeling, and prediction (pp. 18-25). Springer, Berlim, Heidelberg.

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

[9] Hawkins, J. B., Brownstein, J. S., Tuli, G., Runels, T., Broecker, K., Nsoesie, E. O., ... & Greaves, F. (2015). Medindo a qualidade do atendimento percebida pelo paciente em hospitais dos EUA usando o Twitter. BMJ Qual Saf, bmjqs- 2015.

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

[11] Blei, D. M., Ng, A. Y., & Jordan, M. I. (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 classificação de sentenças. arXiv preprint arXiv:1408.5882.

[14] Yih, W. T., Toutanova, K., Platt, J. C., & Meek, C. (2011, junho). Projeções de aprendizagem discriminatórias para medidas de similaridade de texto. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning (pp. 247-256). Association for Computational Linguistics (Associação para Linguística Computacional).

[15] Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014, abril). Aprendendo representações semânticas usando redes neurais convolucionais para busca na web. Em Anais da 23ª Conferência Internacional na World Wide Web (pp. 373-374). ACM.

[16] Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). Uma rede neural convolucional para modelagem de frases. arXiv preprint 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 idiomas, visualização e interfaces (pp. 63-70).