Rastreador de sintomas da COVID19

Ticiana L. Coelho da Silva[1], **Marianna Gonçalves F. Ferreira**[1], **Regis Pires Magalhães**[1], **José Antônio F. de Macêdo**[1], **Natanael da Silva Araújo**[1]

[1]Insight Data Science Lab – Universidade Federal do Ceará (UFC)
Caixa Postal 60.440-900 – Fortaleza – CE – Brasil

{ticianalc, marianna, regis, jose.macedo, natanaelsilva}@insightlab.ufc.br

Abstract. *The pandemic caused by coronavirus has fueled the need for technological solutions capable of capturing and monitoring data in an automatic, agile and secure way. The coronavirus on-call service system made available to the population of the Ceará State has automatic symptom recognition technology through Natural Language Processing (NLP). The tracker proposed in this work, called Sintomatic, is a neural network that processes texts, capturing symptoms in messages exchanged between the citizen of Ceará and the nurse/doctor at the Plantão Coronavirus. In addition, Sintomatic identifies and captures mental health behaviors, such as: anxiety, distress and trends in depression.*

Resumo. *A pandemia causada por coronavírus fomentou a necessidade de soluções tecnológicas capazes de capturar e monitorar dados de forma automática, ágil e segura. A Plataforma de Atendimento Plantão Coronavírus disponibilizada para a população do Estado do Ceará possui tecnologia de reconhecimento automático de sintomas por meio do Processamento de Linguagem Natural (NLP). O rastreador proposto neste trabalho, chamado de Sintomatic, é uma rede neural que processa textos, capturando sintomas em mensagens trocadas entre o cidadão cearense e o enfermeiro/médico no Plantão Coronavírus. Além disso, o Sintomatic identifica e captura comportamentos de saúde mental, como: ansiedade, angústia e tendências de depressão.*

1. Introdução

Diante do cenário causado pela pandemia por coronavírus e o acometimento no Estado do Ceará desde março de 2020, surgiu a demanda do desenvolvimento soluções tecnológicas que fossem capazes de capturar e monitorar dados de forma automática, ágil e segura, pois pouco sabia-se sobre o comportamento a e evolução do vírus.

Gestores de saúde e tomadores de decisão necessitavam de dados para mitigar um período de desconhecimento e incertezas. Conhecer os padrões da doença era crucial para elaborar protocolos de saúde eficientes. Para isso, era preciso reconhecer sintomas e comportamentos de saúde mental da população acometida. Uma das soluções desenvolvidas e disponibilizadas para a população no Estado do Ceará foi o Plantão Coronavírus, uma plataforma com mecanismos de triagem que, no primeiro momento utiliza um *chatbot* para interagir com o paciente a fim de classificar seu estado de saúde em uma das três categorias: verde, amarelo e vermelho, sendo o nível de criticidade da saúde do paciente leve, moderada ou grave, respectivamente. Após a triagem com o chat de teleatendimento

do Ceará, ele pode ser encaminhado para uma tele consulta com um profissional de saúde, a depender do seu quadro clínico.

Mesmo de posse dos dados das conversas entre o paciente e o profissional de saúde por meio do Plantão Coronavírus, seria inviável para a Secretaria de Saúde do Estado do Ceará rastrear os sinais da doença manualmente por meio da leitura de milhares de relatos. Dessa forma, era necessário uma solução automatizada e inteligente para reconhecimento dos padrões da COVID19.

Este trabalho mapeou a identificação de sintomas em texto como um problema de reconhecimento de entidade (em inglês, *Named Entity Recognition* – NER). NER corresponde à capacidade de identificar as entidades nomeadas nos documentos e rotulá-las em classes definidas de acordo com o tipo de entidade [da Silva et al. 2019]. De forma geral, o robô de captura de sintomas possui uma rede neural que é capaz de reconhecer entidades. Neste trabalho, uma entidade é um sintoma.

O mecanismo de captura de sintomas perpassa por todo o processo de triagem com o *chatbot*, até o tele atendimento com o profissional de saúde. O robô de captura de sintomas, chamado de Sintomatic, é a principal contribuição deste trabalho. Sintomatic é uma tecnologia que consome os dados da plataforma Plantão Coronavírus, e então é capaz de processar e identificar os sintomas contidos nos textos em linguagem natural, utilizando Processamento de Linguagem Natural (PLN), tecnologia largamente utilizada para ajudar computadores a entender a linguagem do ser humano. O link [1] apresenta uma breve demonstração do Sintomatic.

Este tipo de inteligência foi essencial para identificar padrões de sinais da doença, bem como novos sintomas ou sintomas raros, que ainda não haviam sido mapeados pelos profissionais de saúde, e, dessa forma, acompanhar a evolução dos achados da COVID19 ao longo dos dias.

O processo de reconhecimento de entidades foi realizado completamente automático, sendo destacado como um diferencial frente aos trabalhos relacionados [Tarcar et al. 2020] que apresentou F1 de 78,5% e [Neumann et al. 2019] que apresentou 84,94% de F1 para o modelo de descoberta de sintomas [2], enquanto o Sintomatic tem F1 igual a 85,66%.

Nas seções seguintes, serão abordadas a metodologia usada na construção do Sintomatic e os cenários de demonstração. E por fim, a conclusão deste artigo.

2. Sintomatic

O Sintomatic é um modelo computacional, que foi desenvolvido com o objetivo de auxiliar a Secretaria de Saúde do Estado do Ceará no acompanhamento dos pacientes que buscavam algum tipo de serviço de saúde, bem como na descoberta de novos sintomas presentes em vítimas do coronavírus, sejam esses mais frequentes ou raros. Devido à possível mutação do vírus e consequente aparecimento de novas ocorrências de sintomas, como foi o caso da anosmia, tornando-se frequente após um certo período da pandemia em pacientes positivo para COVID19, este modelo proporcionou grandes ganhos no entendimento da doença pela sua capacidade de reconhecer novos padrões.

[1] https://bit.ly/sintomatic

[2] https://allenai.github.io/scispacy/

O Sintomatic é uma rede neural que processa textos em Linguagem Natural, capaz de identificar sintomas a partir de mensagens trocadas entre o *chatbot* e o paciente, bem como reconhecer novos padrões da doença anteriormente inexistente ou despercebidos. Esse tipo de inteligência pode ser perfeitamente treinado para reconhecer e capturar outras classes de palavras em qualquer contexto desejado.

A detecção de sintomas no idioma português foi um desafio, pois, até o momento, não havia de forma pública nenhum modelo capaz de realizar essa tarefa, de acordo com o conhecimento dos autores. O robô desenvolvido foi treinado através de um processo de aprendizado conhecido como *Transfer Learning* [Pan and Yang 2009], ou em português, aprendizado por transferência.

A técnica de aprendizagem por transferência utiliza o conhecimento adquirido ao resolver um problema e aplicá-lo em outro problema diferente, porém relacionado, permitindo progresso rápido e desempenho aprimorado ao modelar a segunda tarefa. Em outras palavras, a transferência de aprendizado é a melhoria do aprendizado em uma nova tarefa através da transferência de conhecimento de uma tarefa relacionada que já foi aprendida.

Para treinar o Sintomatic foi utilizado o scispaCy, um pacote Python que contém modelos de spaCy [Honnibal and Montani 2017] para processar textos biomédicos, científicos ou clínicos. Em particular, há um tokenizador personalizado que adiciona regras de tokenização baseando-se em regras do spaCy, um etiquetador POS e analisador sintático treinado em dados biomédicos e um modelo de detecção de extensão de entidade. Separadamente, também existem modelos NER para tarefas mais específicas. Para este trabalho o modelo utilizado foi o en_ner_bc5cdr_md do SciSpacy, em um processo de *transfer leaning* para treinar um novo modelo de reconhecimento e captura de sintomas em português.

A primeira etapa do processo de treino do rastreador foi traduzir os textos que inicialmente estavam em língua portuguesa para o idioma inglês. Em seguida, inserir como parâmetro de entrada cada texto ao modelo do scispacy, analisar o resultado gerado por este modelo, e logo após traduzir os sintomas capturados pelo modelo do scispacy em inglês para português. O conjunto de treinamento para o Sintomatic (novo modelo em português), é composto do texto original e os sintomas capturados pelo modelo do scispacy em português. Esse processo foi executado de forma contínua até que a função de erro da rede se estabilizasse. Ao final, foi possível atingir para o Sintomatic, *F1-score* de 85.66, o que é competitivo se comparado ao modelo em inglês, que tem *F1-score* igual a 85.02. A Figura 1 ilustra as etapas desse processo.

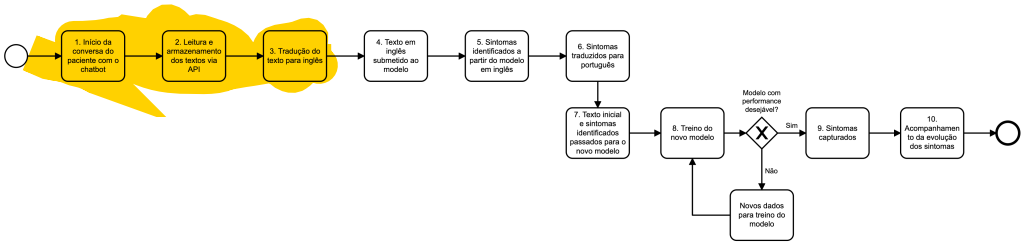


Figura 1. Fluxo dos dados

Para subdividir os sintomas capturados pelo robô, foi criada uma API com três chamadas principais, a fim de retornar informações sobre os resultados mais constantes e

raros, conforme exposto na Figura 2.

Endpoint	Metodo HTTP	Descrição e Exemplo
getRanking	GET	Usado para buscar os dez sintomas mais frequentes dentro do intervalo de datas passado como parâmetro
getRankingPercentage	GET	Retorna uma lista com os dez sintomas mais frequentes e sua representação percentual
getRankingRare	GET	Retorna os dez sintomas menos relatados pelos pacientes, considerados raros
getEvolution	GET	Retorna a evolução dos sintomas mais comuns dentro do intervalo de data passado como parâmetro
dailySymptoms	POST	Usado para para inserir os sintomas no banco
getRankingSymptoms	GET	Retorna todos os sintomas que foram respondidos pelos pacientes como "sim" e a respectiva categoria deste paciente (verde, amarela e vermelha)
getRankingHealthProblems	GET	Retorna todos os problemas de saúde que foram respondidos pelos pacientes como "sim" e a respectiva categoria deste paciente (verde, amarela e vermelha)
getTotalPatients	GET	Retorna o total de pacientes que foram triados e o total para cada categoria (verde, vermelha e amarela)

Figura 2. Descrição da API

Um diferencial do Sintomatic é a ausência da necessidade de classificação manual realizada por um humano para reconhecimento de entidades. Em um cenário onde havia vasta quantidade de dados e pouco tempo para processar essas informações, o ganho com a otimização dessa etapa de treino foi crucial no apoio a tomada de decisão.

Outro quesito inovador promovido pelo robô de captura, foi a capacidade que essa tecnologia desenvolveu de aprender a reconhecer comportamentos de saúde mental como sintomas recorrentes em pacientes suspeitos ou não para COVID19. A partir desta contribuição profissionais de saúde e respectivos órgãos competentes, podem valer-se de tal dado para elaborar e promover políticas públicas com o propósito de assistir a essas pessoas que são acometidas por problemas que ultrapassam a esfera epidemiológica.

Atualmente, o Sintomatic é utilizado na plataforma de Tele Atendimento do Estado do Ceará, onde desempenha papel pioneiro na área da saúde.

3. Cenários de Demonstração

Em um momento de grandes transformações ocasionados pela pandemia por COVID19, surgiu a necessidade de escalar um serviço de saúde de forma rápida e segura, tanto para pacientes como para profissionais de saúde. A partir desse cenário, foi disponibilizado para a população do Estado do Ceará um serviço de Tele Atendimento gratuito, onde o paciente inicialmente trocava mensagens com um robô, era triado de acordo com seus sintomas e, posteriormente, encaminhado para uma consulta com um profissional de saúde.

Todo esse ciclo de integração com o paciente registrado por meio de textos é passado ao modelo Sintomatic para que este possa detectar sintomas em todas as etapas do atendimento.

A Figura 3 exemplifica parte de uma conversa com um paciente anônimo:

Figura 3. Trecho da conversa entre o paciente e o *chatbot*

Para o acompanhamento dos dados capturados pelo robô Sintomatic e monitoramento das demais informações sobre a pândemia, foi desenvolvido o Boletim Digital COVID-19 do Ceará, solução tecnológica construída por cientistas de dados onde é feito todo o processo de mineração do dado bruto até sua exposição em painéis gráficos acompanhados de textos explicativos à respeito de cada uma das análises abaixo:

- número de pacientes atendidos;
- sintomas mais frequentes e raros;
- evolução dos sintomas por semana epidemiológica;
- sintomas ao longo do tempo.

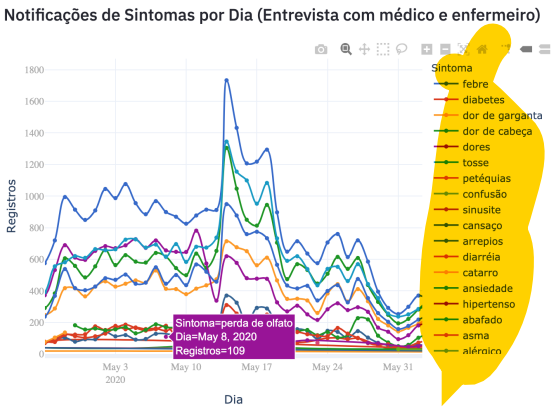
A Figura 4 ilustrada no próximo capítulo desta demonstração expõe a evolução dos sintomas em uma série temporal. Através dessa imagem pode-se identificar a detecção de um novo sintoma no dia oito de maio, perda de olfato. Este sintoma apareceu e tornou-se bastante característico da COVID19 após um certo período de tempo. Ainda sobre a série temporal, é possível visualizar que a frequência de cada sintoma é sazonal durante o período analisado. Comportamentos de saúde mental, como ansiedade também podem ser observados dentre os sintomas desse gráfico.

Todo o processo de desenvolvimento e treino do Sintomatic foi realizado em um curto intervalo de tempo e com poucos recursos de mão de obra, pois, dada a atipicidade e urgência da situação, este foi um projeto que precisou ser desenvolvido sem a tradicional estruturação prévia.

Demais estudos também foram realizados com a finalidade de subsidiar a tomada de decisão guiada por dados. Todos os analíticos estão disponíveis no Boletim Digital COVID19 no Ceará.

4. Conclusão

Nessa demonstração, é proposto o uso de um modelo de aprendizado de máquinas para identificar sintomas e comportamentos mentais alterados na população do Estado do Ceará durante a crise causada por coronavírus, chamado Sintomatic. Ao longo de quatro meses de pandemia no Estado, diversos sintomas foram visualizados. É possível verificar, na imagem abaixo, que a manifestação desses sintomas variam consideravelmente em relação ao tempo, assim como novas ocorrências também foram identificadas ao longo do período analisado.



Fonte: Ceará Tele Atendimento

Figura 4. Sintomas ao longo dos dias

Além de reconhecer novos padrões de sintomas causados por SARS-COV-2, uma das principais contribuições deste trabalho é identificar comportamentos psicológicos alterados, como: ansiedade, angústia e tristeza em pacientes positivos ou não para COVID19.

Diante dessa informação, a Secretaria de Saúde do Estado do Ceará, ou qualquer outro órgão que faça uso dessa tecnologia, pode desenvolver políticas com o propósito de acompanhar essas pessoas em um quadro clínico que acomete não apenas sua saúde fisiológico, como também emocional.

Referências

da Silva, T. L. C., Magalhães, R. P., de Macêdo, J. A., Araújo, D., Araújo, N., de Melo, V., Olímpio, P., Rego, P. A., and Neto, A. V. L. (2019). Improving named entity recognition using deep learning with human in the loop. In *EDBT*, pages 594–597.

Honnibal, M. and Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1).

Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). Scispacy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Tarcar, A. K., Tiwari, A., Rao, D., Dhaimodker, V. N., Rebelo, P., and Desai, R. (2020). Healthcare ner models using language model pretraining. In *HSDM@ WSDM*, pages 12–18.