Guia para o Processo de Anotação

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

<omitido para revisão>

Sumário

1. Detalhamento da anotação

Para o processo de anotação serão definidos alguns direcionamentos para os anotadores, isto é importante pois garantirá que o processo de anotação terá um mesmo critério por parte dos anotadores.

1.1. Sobre as classes

Abaixo se encontram o código da classe, a classe e sua definição.

1 = Fake news: Tweet sobre a covid-19/coronavírus com características de notícia, mas que não há notícias oficiais corroborando com o seu conteúdo.

Tuíte que descredibiliza uma informação cientificamente comprovada.

Ex.: "Tomo ivermectina, trabalho na rua e graças a Deus não tive covid19.", fere a comprovação científica de que o antiparasitário não previne a Covid-19

Ex.: "Vai deixar mesmo algum parente seu tomar a VaChina da Sinovac?", fere a comprovação científica de que a vacina Coronavac possui eficácia comprovada.

0 = Neutro: Tweet, apesar de envolver a covid-19/coronavírus não se trata de uma informação ou notícia. Por exemplo, tweets que expressam opinião (exemplos e/ou histórias que não ferem um conhecimento com comprovação científica. Tuítes que não são passíveis de verificação).

-1 = News: Tweet sobre a covid-19/coronavírus com características de notícia e que há notícias oficiais corroborando com o seu conteúdo.

Opção dúvida: O anotador deverá marcar essa opção quando sentir dúvida quanto a classificação do tweet. Mesmo que marque essa opção, o tweet ainda deverá ser classificado em uma das classes.

1.2. Sobre o pertencimento dos tweets para com as respectivas classes (Fake News, Neutro, News)

1. Com o objetivo de possibilitar uma anotação com um mínimo de viés, definiu-se uma sequência de estratégias para verificar a legitimidade do conteúdo dos tweets:
2. Confirmação da notícia em veículos de divulgação científica Nacionais/Internacionais, especificamente, da área de saúde (possuir fator de impacto)
3. Confirmação da notícia em veículos oficiais do Governo (Ministério da Saúde, ANVISA ...);
4. Confirmação da notícia em veículos de comunicação que compõem o consórcio de imprensa para levantamento de dados sobre a Covid-19;
5. Confirmação da notícia em veículos de comunicação que compõem a Associação Nacional de Jornais ANJ;
6. Confirmação da notícia na Organização Mundial da Saúde;
7. Confirmação de notícias em veículos de comunicação que fazem parte da lista da Forbes <https://www.forbes.com/sites/berlinschoolofcreativeleadership/2017/02/01/10-journalis

m-brands-where-you-will-find-real-facts-rather-than-alternative-facts/?sh=3380f522e9b5>

8. Verificar se pelo menos 2 dos veículos confirmam a legitimidade do conteúdo do tweet.

1.3. Sobre a anotação

Dois anotadores serão responsáveis por tratar da classificação manual do grupo de treinamento que será utilizado nos algoritmos de aprendizado de máquina. Para que se tenha sucesso em tal atividade, os anotadores deverão adotar os seguintes procedimentos:

a. O grupo de treinamento será composto por uma quantidade x de tweets;
b. Todos os avaliadores receberão o mesmo pacote de tweets;
c. Os anotadores deverão avaliar individualmente cada tweet e classificá-lo;
d. Os anotadores não poderão realizar anotações por períodos superiores a 25 minutos com intervalo mínimo de 5 min;
e. A qualquer momento dentro do período delimitado pode ser selecionado a opção dúvida;
f. Se dentro do tempo delimitado o avaliador não conseguir classificar o tweet, ele deverá marcar a opção dúvida e escolher a classe que ele considera mais adequada para o tweet;
g. Um dos anotadores deve conduzir a anotação de forma crescente (iniciando do primeiro tweet até o último) enquanto o segundo anotador conduz de forma decrescente (iniciando do último tweet até o primeiro);
I. Um terceiro anotador será acionado para resolver todas divergências de anotação, isto é, os casos em que não houve concordância entre os anotadores primários.

1.4. Avaliação da concordância entre os avaliadores

Para avaliar o grau de concordância dos anotadores primários, será utilizada a métrica Cohen's Kappa
<https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat00365.pub2>

2. Tratamento de tweets

Para a formação do grupo de treinamento é necessário fazer uso de um crawler/robô, da forma que foi implementado ele busca os tweets contentando as expressões "covid", "covid-19" e "coronavírus", é armazenado a data do tweet, a localização do usuário que postou e o tweet em si. Esse tweet é recuperado da mesma forma que foi postado, ou seja, ele vai conter expressões linguísticas, abreviações, emojis e urls. Como essas expressões não fazem parte da língua portuguesa é necessário que se faça um tratamento dos tweets. Esse processo é importante pois ajudará os anotadores em sua tarefa e também evitará erros na fase de aprendizado de máquina.

2.1. Abreviações e expressões linguísticas

Por se tratar da natureza da comunicação na Internet e por conter uma quantidade de caracteres limitados por tweet no Twitter é comum encontrarmos expressões únicas da

internet, essas podem ser tanto abreviações de palavras ou expressões como também podem ter um significado totalmente novo.
Para o tratamento de tal expressão pode ser necessário fazer um apanhado das mesmas, podendo ser através de um formulário contendo a expressão e seu significado para que seja substituído no tweet original.

2.2. URLs

É comum no Twitter os tweets conterem URLs que se referem a outro post ou a um site externo. Como não se é praticável abrir cada link para ver seu conteúdo, se faz necessário que os mesmos sejam desconsiderados dos tweets originais.

2.3. Emojis

Os emojis podem ser definidos como caracteres especiais que representam sentimentos das mais variadas formas, fazem parte da cultura da Internet e são muito comuns dentro da comunicação da mesma.
Para o tratamento de tais expressões pode-se avaliar se os mesmos podem ser interpretados através do seu código (significado do caractere para a máquina) e considerá-los. Uma outra opção é a remoção dos mesmos dos tweets, mesmo que isto represente uma perda de informação.