

Grupo 6 (?)

Relatório do projeto final – Text Mining

Nome do projeto

Análise de Sentimentos e Modelação de Tópicos de Tweets relacionados com a vacina AstraZeneca

Âmbito do projeto

Trabalho realizado no âmbito da Unidade Curricular de Text Mining

Constituição do grupo

Guilherme Correia – DSA – Nº120088

Tiago Vaz – TDIA – Nº120055

Vasco Silva – TDIA – Nº120050

Contents

[1. Introdução 3](#_Toc185453119)

[2. Estado da arte 4](#_Toc185453120)

[3. Metodologia 5](#_Toc185453121)

[3.1 Extração de dados 5](#_Toc185453122)

[3.2 Pré-processamento de dados 5](#_Toc185453123)

[3.3 Análise de Tópicos 6](#_Toc185453124)

[4. Referências 7](#_Toc185453125)

# Introdução

Nos últimos anos, as redes sociais têm desempenhado cada vez mais um papel importantíssimo no que toca ao desenvolvimento da opinião pública, especialmente quando passamos por uma crise global, nomeadamente a pandemia de COVID-19. O Twitter é uma das plataformas mais influenciais tendo comunicação praticamente instantânea e com isso acabou por se tornar numa via fundamental para discussões sobre a saúde pública, incluindo o debate sobre as vacinas contra a COVID-19. Para este trabalho, dentro das diversas vacinas que foram criadas, a AstraZeneca, é a que nos interessa. Criada pela Universidade de Oxford e a farmacêutica AstraZeneca, a vacina gerou uma grande quantidade de conversas, tanto positivas como negativas, sendo um reflexo das várias controvérsias em que esteve envolvida, sobre a sua segurança, eficácia e efeitos secundários.

O objetivo deste projeto é fazer uma análise de sentimento e modelação de tópicos em tweets relacionados com a vacina da AstraZeneca, com o propósito de perceber melhor as opiniões do público, enquanto aplicamos os vários conhecimentos adquiridos durante o semestre na Unidade Curricular de Text Mining. Queremos usar particularmente as várias técnicas aprendidas como a análise de sentimentos, modelação de tópicos, pré-processamento de dados (isto inclui várias áreas), entre vários outros pontos. Isto tudo depois será interligado para perceber se o dataset que utilizámos funciona corretamente ao aplicar vários modelos e as suas nuances.

# Estado da arte

Passados quatro anos depois da Organização Mundial de Saúde ter declarado a SARS-CoV-2 uma emergência de saúde pública, foram feitos vários estudos acerca do tema, em especial da opinião das pessoas em relação às vacinas. Uma das plataformas em que se pode encontrar mais informações e opiniões sobre o tópico é o “Twitter”.

O primeiro estudo[[1]](#footnote-2), publicado no dia 9 de agosto de 2021 pela Universidade de Medicina de Zagreb, utilizou o “Twitter” para encontrar cerca de setecentos mil “tweets” com opiniões sobre as várias vacinas, desde o dia 1 de dezembro de 2020 ao dia 31 de março de 2021. Os resultados demonstraram que a opinião sobre as vacinas “Pfizer” e “Moderna” manteve-se estável entre os quatro meses em questão, contudo as opiniões sobre a vacina da “AstraZeneca” diminuíram drasticamente.

O segundo estudo[[2]](#footnote-3) foi publicado em julho de 2021 no “Indonesian Journal of Electrical Engineering and Computer Science” e é parecido com o anterior, utilizando a API do Twitter foram capazes de extrair dez mil tweets para cada uma das vacinas, utilizando hashtags relativas a cada uma, “AstraZeneca”, “Pfizer” e “Moderna”. A partir da análise podemos perceber que a “Pfizer” demonstra mais confiança que as outras duas, cerca de 47.29% dos “tweets” são positivos, 37.5% negativos e 15.21% neutros. A “Moderna” tem 46.16% dos tweets positivos, 40.71% negativos e 13.13% neutros. Por último, e com mais importância para nós, a “AstraZeneca” foi a que demonstrou menos confiança no total, com 40.08% dos tweets positivos, 40.06% negativos e 13.86% neutros.

Portanto, em relação ao nosso dataset que se foca especialmente nas vacinas da “AstraZeneca”, a opinião geral que resultou dos estudos, demonstra que comparando com as restantes, estas vacinas têm geralmente um nível de confiança muito mais baixo, tendo até diminuído no espaço de tempo logo depois do “fim” do primeiro ano de COVID-19.

# Metodologia

## 3.1 Extração de dados

Em relação à metodologia utilizada no projeto começámos por procurar por um dataset para analisar acabando por chegar a este[[3]](#footnote-4) sobre as vacinas da AstraZeneca e o Twitter. Este dataset é uma compilação de tweets reais retirados do website através da API pública do Twitter, utilizando palavras-chave específicas e hashtags relacionados com as entidades e tópicos em questão. No caso dos dados em si, tem cerca de 1500 tweets, cada um com subjetividade (valor numérico que representa o quão subjetivo é o tweet), polaridade (valor número do sentimento que o tweet demonstra) e o objetivo que é saber se é positivo, negativo ou neutro. O único problema é não ter um intervalo temporal para a extração dos tweets.

## 3.2 Pré-processamento de dados

Nesta secção começámos por introduzir ler o dataset no “notebook” utilizando a biblioteca “pandas”. Depois fizemos a limpeza inicial, retirar as colunas desnecessárias como o “id”, “subjetividade“ e “polaridade”, perceber se existe algum valor nulo, retirar duplicados e guardar a coluna “Target”, esta será muito importante depois para a análise e modelamento dos tópicos. Também normalizámos o texto presente nos tweets através de “regex”.

Depois fizemos a remoção de “stopwords” através da biblioteca “NLTK”, primeiro fazemos a tokenização e depois removemos as “stopwords” que estão presentes na lista da língua inglesa. Por último fazemos a lematização através do “WordNetLemmatizer”, que consiste em reduzir as palavras à sua forma base, baseando-se na biblioteca “WordNet”. Depois reconstruímos os tweets com as palavras lematizadas e criamos um vocabulário único com essas mesmas palavras. Este processo é muito importante para padronizar as palavras e criar um vocabulário básico, essencial para a análise de sentimentos, modelagem de tópicos e por fim treinar modelos de Machine Learning. Escolhemos usar a lematização em vez da stemmatização por ser uma alternativa mais sofisticada que em vez de remover apenas os sufixos e chegar à raiz da palavra, o WordNetLemmatizer usa o contexto gramatical da palavra para produzir um resultado mais preciso, assim as palavras não perdem o seu significado.

## 3.3 Análise de Tópicos

# Referências

* Marcec, R., & Likic, R. (2022). Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. *Postgraduate Medical Journal*, *98*(1161), 544–550. <https://doi.org/10.1136/postgradmedj-2021-140685>
* Shamrat, F. M. J. M., Chakraborty, S., Imran, M. M., Muna, J. N., Billah, M. M., Das, P., & Rahman, M. O. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, *23*(1), 463–470. <https://doi.org/10.11591/ijeecs.v23.i1.pp463-470>
* Rafael, G. (2021). *Twitter AstraZeneca AntiCovid [Dataset]*. Kaggle. Retirado de <https://www.kaggle.com/datasets/gallo33henrique/twitter-astrazeneca-anticovid>

1. Marcec, R., & Likic, R. (2022). Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. *Postgraduate Medical Journal*, *98*(1161), 544–550. <https://doi.org/10.1136/postgradmedj-2021-140685> [↑](#footnote-ref-2)
2. Shamrat, F. M. J. M., Chakraborty, S., Imran, M. M., Muna, J. N., Billah, M. M., Das, P., & Rahman, M. O. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, *23*(1), 463–470. <https://doi.org/10.11591/ijeecs.v23.i1.pp463-470> [↑](#footnote-ref-3)
3. Henrique, G. (2021). *Twitter AstraZeneca AntiCovid [Dataset]*. Kaggle. Retrieved from <https://www.kaggle.com/datasets/gallo33henrique/twitter-astrazeneca-anticovid> [↑](#footnote-ref-4)