

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

Índice

[1. Introdução 3](#_Toc185611970)

[2. Estado da arte 4](#_Toc185611971)

[3. Metodologia 5](#_Toc185611972)

[3.1 Extração de dados 5](#_Toc185611973)

[3.2 Pré-processamento de dados 5](#_Toc185611974)

[3.3 Análise de Tópicos 6](#_Toc185611975)

[3.4 Classificação de sentimentos 6](#_Toc185611976)

[4. Análises e conclusões 7](#_Toc185611977)

[4.1 Análise de tópicos 7](#_Toc185611978)

[4.2 Classificação de sentimentos 8](#_Toc185611979)

[4.3 Conclusões finais 11](#_Toc185611980)

[5. Referências 12](#_Toc185611981)

Índice de imagens

[Figura 1 – Matriz de confusão da análise de tópicos](#_Toc185612017) 7

[Figura 2 – Representação interativa do LDA 8](#_Toc185612018)

[Figura 3 – Relatório de classificação do Naive Bayes 8](#_Toc185612017)

[Figura 4 – Matriz de confusão do Naive Bayes 9](#_Toc185612018)

[Figura 5 – Relatório de classificação da Regressão Logística 9](#_Toc185612017)

[Figura 6 – Matriz de confusão da Regressão Logística 9](#_Toc185612018)

[Figura 7 – Matriz de confusão do analisador de sentimento VADER 10](#_Toc185612017)

# 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

Nesta secção foi feita uma análise dos tweets através de diversos tópicos criados por modelos treinados através do nosso dataset. Primeiro utilizamos o TF-IDF, um modelo capaz de transformar cada tweet num vetor com a representação numérica de cada palavra, frequência da mesma e a sua raridade, acabando por nos dar uma matriz em que cada linha representa um tweet e cada coluna uma única palavra. Depois aplicamos o modelo NMF (“Non-negative matrix factorization”), aqui o objetivo é criar duas matrizes de forma a identificar os tópicos principais de cada tweet. A matriz “W” representa os pesos dos tweets em relação aos tópicos, a matriz “H” representa o peso das palavras em relação aos tópicos, daqui retiramos três tópicos. Por fim, agora somos capazes de extrair as palavras presentes em cada tópico. No dataframe acabamos por adicionar uma nova coluna com o tópico dominante para cada tweet e comparar com a coluna “target”, sendo que 0 é negativo, 1 é positivo e 2 é neutro. Para fazer esta comparação utilizamos uma matriz de confusão.

Na segunda parte aplicamos um modelo diferente que utiliza também o TF-IDF, mas depois passa para um “Bag of Words”. Aqui já recorremos a uma biblioteca diferente também para usar o LDA (“Latent Dirichlet Allocation”), primeiro criamos um vocabulário novo e depois transformamos cada tweet numa representeação número no formato “Bag of Words”. Por fim treinamos o modelo LDA para identificar os três tópicos mais importantes, ajustando os pesos das palavras em cada tópico após 10 iterações. Os tópicos criados são listados com as palavras mais representadas e os seus pesos, permitindo identificar os temas principais discutidos nos tweets. Acabando por utilizar o “PyLDAvis” para criar uma visualização interativa, onde cada tópico é representado como uma esfera. O tamanho das esferas reflete a relevância dos tópicos, e as palavras-chave ajudam a interpretá-los.

## 3.4 Classificação de sentimentos

Para esta secção o objetivo é treinar vários modelos para compreender se as palavras presentes em cada tweet são capazes de nos levar ao “target”. Primeiro recorremos a dois modelos diferentes, mas com aplicações muito parecidas: Naive Bayes Multinominal e Regressão Logística. Ambos são modelos de classificação capazes de trabalhar com várias classes como sentimentos e atribuem probabilidades a cada antes da previsão final, contudo enquanto o Naive Bayes assume que cada palavra é condicionalmente independente, a Regressão Logística baseia-se numa função logarítmica para modelar a relação entre as características e probabilidades. O Naive Bayes também é um modelo simples e ideal para conjuntos de dados pequenos e a Regressão Logística é mais robusta, mas muito flexível.

Em relação à aplicação de cada modelo, começamos por dividir o dataset em conjuntos de treino e teste, treinamos cada modelo de formas parecidas para eles nos devolverem previsões para os tweets nos conjuntos de teste. Por último avaliamos os dois modelos com recurso a um relatório de classificação e matrizes de confusão.

Depois destes dois modelos decidimos utilizar um diferente, o analisador de sentimentos VADER (“Valence Aware Dictionary and sEntiment Reasoner”), que avalia a correspondência entre os rótulos reais de sentimento e os rótulos previstos. Primeiro tivemos de criar uma função que lesse cada tweet e lhe atribuísse um valor para cada sentimento, seja ele negativo, positivo ou neutro. É importante realçar que ao contrário dos modelos anteriores, este analisador de sentimentos é binário, logo precisamos de retirar todos os tweets que sejam considerados neutros no “target” já presente desde o início. Por último, depois de acrescentar as previsões ao dataset podemos mais uma vez utilizar a matriz de confusão para analisar os resultados.

# Análises e conclusões

No ponto anterior demonstrámos como foram utilizados e treinados os modelos, nesta secção iremos demonstrar os resultados e as conclusões que conseguimos retirar de cada.

## 4.1 Análise de tópicos

Uma imagem com texto, captura de ecrã, diagrama, design

Descrição gerada automaticamente Em relação à matriz de confusão criada através da análise de tópicos:

Figura 1 - Matriz de confusão da análise de tópicos

Conseguimos observar que os tópicos 2 e 3 são fortemente associados a tweets positivos, por outro lado as classes negativo e neutro apresentam uma correlação menor em comparação com os 3 tópicos. Também é observável que o tópico 1 não apresenta nenhuma relação com nenhum das classes.

Uma imagem com texto, captura de ecrã, diagrama, círculo

Descrição gerada automaticamente No caso da representação interativa criada pelo LDA:

Figura 2 - Representação interativa do LDA

Esta interface é interessante para perceber quais as palavras que têm mais impacto na atribuição de certos sentimentos aos tópicos.

## 4.2 Classificação de sentimentos

Uma imagem com texto, captura de ecrã, Tipo de letra, número

Descrição gerada automaticamente Como foi dito anteriormente, aqui foram utilizados três modelos diferentes, sendo que será mais interessante compara o Naive Bayes com a Regressão Logística e analisar o VADER de forma independente. No caso do Naive Bayes:

Figura 3 - Relatório de classificação do Naive Bayes

Uma imagem com texto, captura de ecrã, diagrama, Retângulo

Descrição gerada automaticamenteUma imagem com texto, captura de ecrã, Tipo de letra, número

Descrição gerada automaticamenteUma imagem com texto, captura de ecrã, diagrama, Retângulo

Descrição gerada automaticamenteRelativamente à Regressão Logística, estes foram os resultados:

Figura 6 - Matriz de confusão da Regressão Logística

Figura 5 - Relatório de classificação da Regressão Logística

Figura 4 - Matriz de confusão do Naive Bayes

Uma imagem com texto, captura de ecrã, diagrama, Retângulo

Descrição gerada automaticamenteConseguimos então perceber que os modelos de Regressão Logística e Naive Bayes apresentam um desempenho geral aceitável, mas com dificuldades devido ao à falta de equilíbrio de classes. A classe "Negativo", com apenas 29 instâncias, é frequentemente mal classificada, sendo confundida com "Neutro" e "Positivo". As classes "Positivo" e "Neutro" têm um desempenho melhor, especialmente na Regressão Logística. Para melhorar os resultados, seria importante equilibrar as classes, ajustando os parâmetros dos modelos ou aplicando técnicas de oversampling ou undersampling. Passando agora para o analisador de sentimentos VADER:

Figura 7 - Matriz de confusão do analisador de sentimentos Vader

A matriz de confusão mostra que o modelo de análise de sentimentos classifica corretamente a maioria dos tweets como "Positivo", mas há uma tendência de confundir as duas classes. A classe "Negativo" é frequentemente classificada como "Positivo", com 202 instâncias de "Negativo" mal classificadas, enquanto a classe "Positivo" também é confundida, com 66 instâncias incorretas. Isso pode ser explicado pela natureza subjetiva e contextualmente variável de sentimentos em alguns tweets. Para além disso este modelo de classificação apenas classifica como "Negativo" ou "Positivo" enquanto o dataset trabalhado apresenta uma terceira classe "Neutro".

## 4.3 Conclusões finais

Para concluir, na análise de tópicos conseguimos perceber, dos tópicos criados pelos modelos, o primeiro tópico não apresenta uma relação muito acentuada com nenhuma das classes de sentimentos, enquanto os outros dois estão constantemente associados as tweets positivos, com as restantes classes a serem muito menos associadas a qualquer um dos tópicos.

Os modelos de Regressão Logística e Naive Bayes apresentam um desempenho geral aceitável, mas são prejudicados pelo desequilíbrio de classes, especialmente a classe "Negativo", que é frequentemente mal classificada. As classes "Positivo" e "Neutro" têm um desempenho melhor, sendo necessário considerar o equilíbrio das classes através de técnicas como oversampling ou undersampling para otimizar os resultados. Além disso, os tópicos 2 e 3 estão fortemente associados a tweets positivos, enquanto a correlação das classes "Negativo" e "Neutro" com os tópicos é mais fraca. O tópico 1 não apresenta qualquer associação significativa com as classes.

A análise de sentimentos revela que, apesar de classificar corretamente a maioria dos tweets como "Positivo", há uma tendência para a classe "Negativo" ser classificada como "Positivo", o que pode ser atribuído à natureza subjetiva e variável dos sentimentos nos tweets. Para além disso, o modelo não contempla a classe "Neutro", que está presente no dataset, limitando a precisão da classificação.

# 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](https://www.kaggle.com/datasets/gallo33henrique/twitter-astrazeneca-anticovid" \t "_new) [↑](#footnote-ref-4)