TEXT-MINE CORONAVIRUS LITERATURE

Goal: use NLP and ML algorithm on the database of research articles to gain useful insight about COVID-19 and how to improve vaccination efforts.

COVID-19 Vaccines Ingredient

➤ Ingredients for all three COVID–19 vaccines currently on the market in US.

MODERNA	PFIZER	JOHNSON & JOHNSON
messenger ribonucleic acid (mRNA), lipids (SM-102, polyethylene glycol [PEG] 2000 dimyristoyl glycerol [DMG], cholesterol, and 1,2-distearoyl-sn-glycero-3-phospho-choline [DSP C]), tromethamine, tromethamine hydrochloride, acetic acid, sodium acetate, and sucrose.	mRNA, lipids ((4-hydroxybutyl)azane-diyl)bis(hexane-6,1-diyl) bis(2-hexyldecanoate), 2 [(polyethylene glycol)-2000]-N,N-ditetrade-cylacetamide, 1,2-Diste-aroyl-sn-glycero-3-phosphocholine, and cholesterol), potassium chloride, monobasic potassium phosphate, sodium chloride, dibasic sodium phosphate dihydrate, and sucrose.	recombinant, replication-incompetentadenovirus type 26expressing the SARS-CoV-2 spike protein, citric acid monohydrate, trisodium citrate dihydrate, ethanol, 2-hydroxy-propyl-β-cyclodextrin (HBCD), polysorbate-80, sodium chloride.

➤ Pfizer-BioNTech and MODERNA are similar vaccines both contain Messenger RNA (mRNA), the main active component will induce the immune response. Classify them together as one group.

➤ The Johnson and Johnson vaccine is little different, it is virus based. The active components are harmless virus.

➤ The bar plot visualize the count of papers contain the ingredients for vaccines. To further improve the model performance, try to use more biomedical terminology for the ingredients.

Visualizing Prevalent Moderna & Pfizer-BioNTech Vaccine ingredients
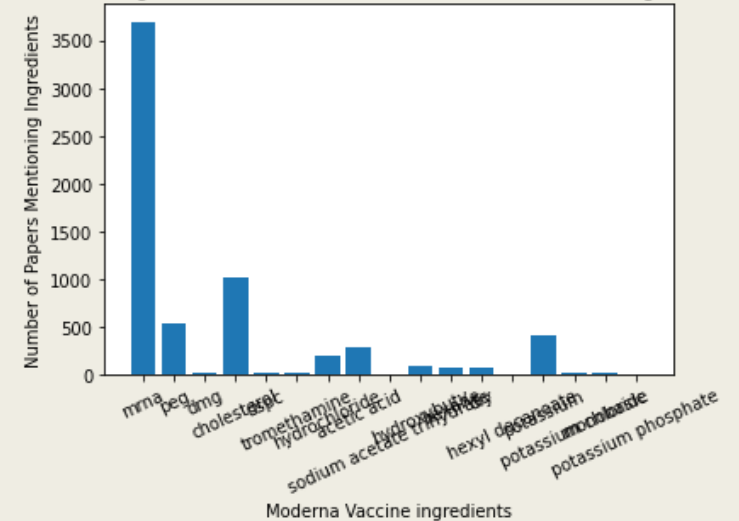
Visualizing Prevalent Janssen Vaccine ingredients
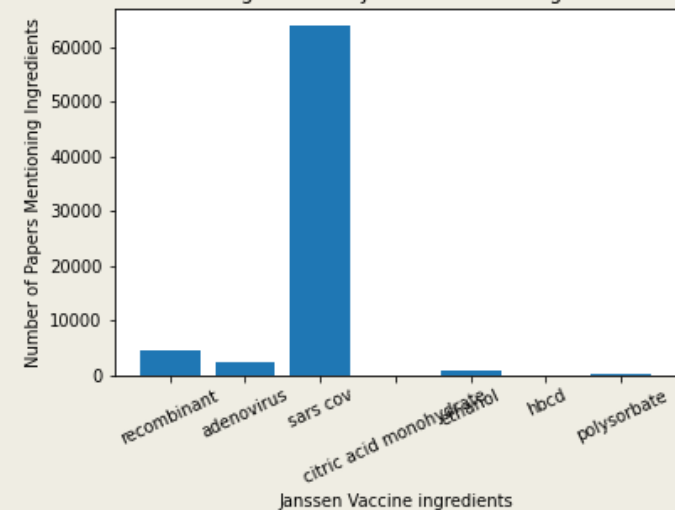
Data Exploration

Define three classes indicate relative vaccine related to the research papers

▶ 'moderna & Pfizer': papers related to the Pfizer-BioNTech and MODERNA vaccines

▶ 'Janssen': papers related to the Johnson and Johnson vaccine

▶ 'other': papers not related to either vaccine

There are 76.81% papers not contain information about vaccine ingredients, so they might not be good resources to study the vaccines. Within the rest 23.19% data set, higher proportion of papers associate to the Johnson and Johnson vaccine. One possible reason could be the virus-based vaccines have a longer history.

Number of papers related to each class varies across the years, use feature 'year' in the modeling process.

Text Exploration

‘moderna & Pfizer’
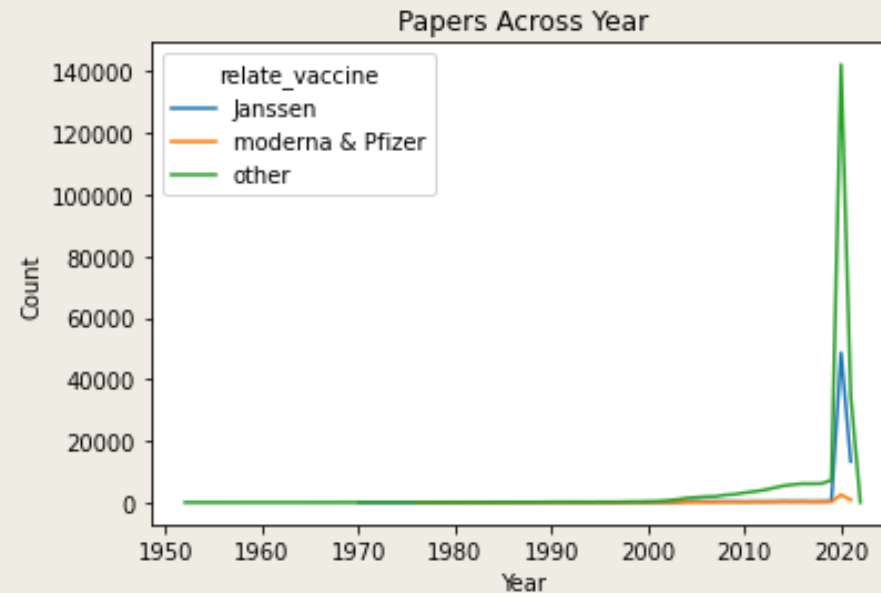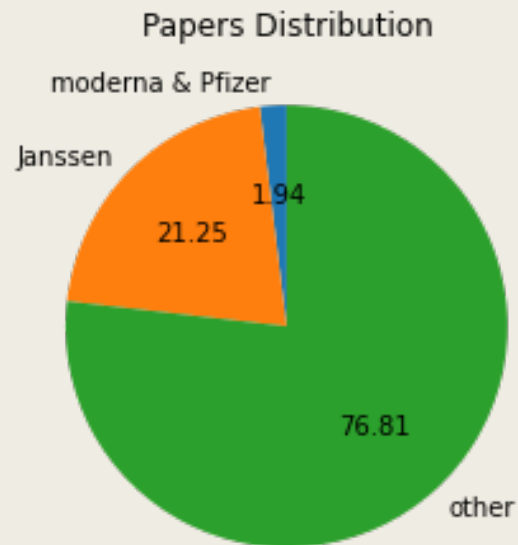
▶ Word cloud visualize the words appear frequently in the corpus of papers related to the vaccine, give better idea of the content.

▶ Words like 'covid', 'patient', 'disease', 'clinical' appears in all plots are neutral word related to the covid-19, which don't contribute much useful information for classification.

▶ For papers contain the frequent words for a certain class, that paper have higher chance also belong to this class. For instance, a paper contains word 'mrna' (main active component of Moderna vaccine), it tends to related to class 'moderna & Pfizer'.
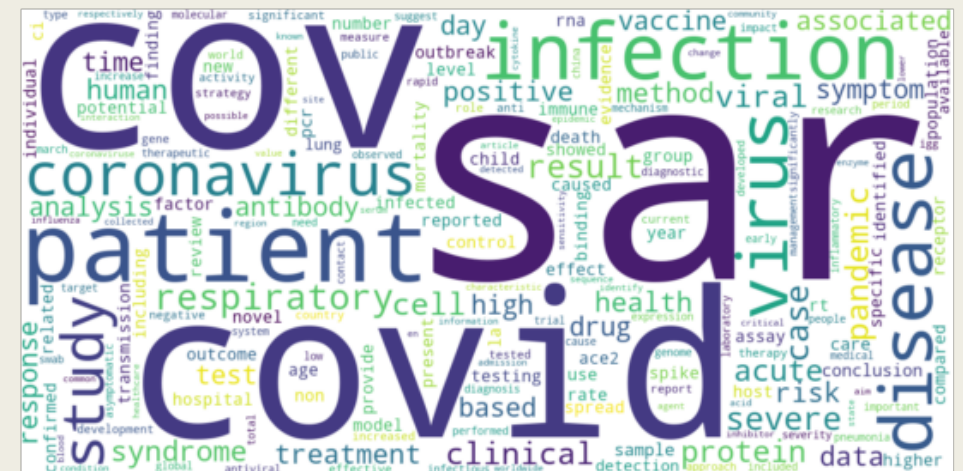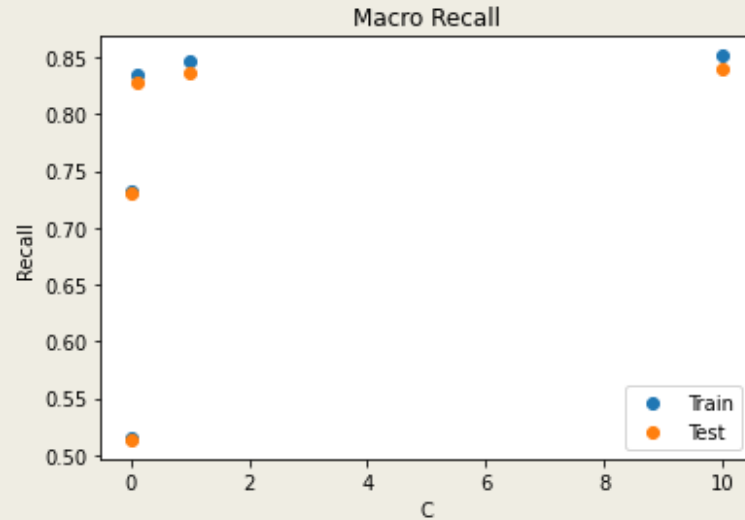
‘other’

‘Janssen’

Multi-Class Classification

Macro Recall

- ► Metric for model performance

  Predict vaccine related papers as non-vaccine related will miss useful information, predict non-vaccine related papers as vaccine related is more accepted. The goal is to minimize false negative and optimize for recall.

  $$R_{macro} = R_{class1} + R_{class2} + R_{class3}$$

  Micro Recall calculates the recall separated by class but not using weights for the aggregation. When the dataset is imbalanced, use Micro Recall as matric since it won't favour the majority class.

- ► Model with highest performance

  Logistic regression with TF-IDF features, l1 penalty with C = 10

	C	Train recall	Test recall
0	0.001	0.515923	0.513618
1	0.010	0.732977	0.730845
2	0.100	0.835318	0.827293
3	1.000	0.846329	0.836466
4	10.000	0.851323	0.840387

True Related Vaccine

Predicted Related Vaccine
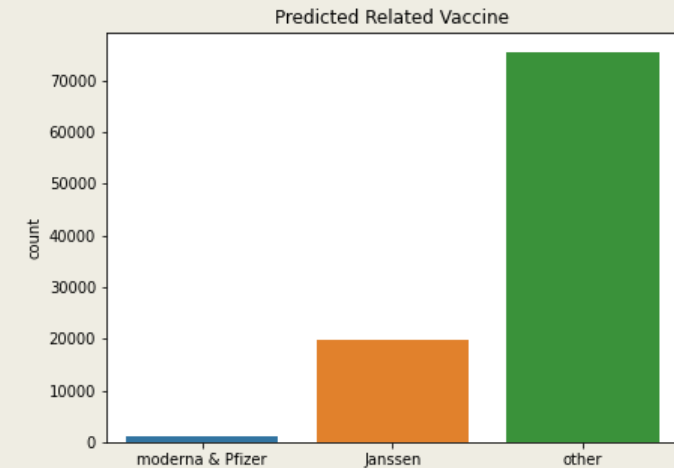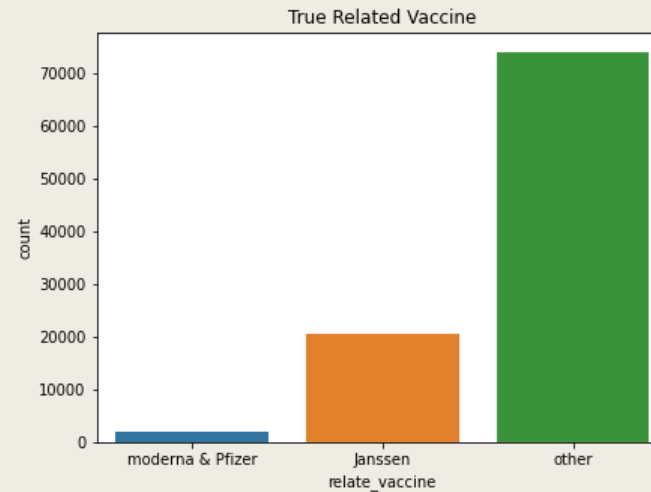
Policy and Guidance

Apply classification model allows scientists to access useful information related to vaccine more efficient. And they can research for the vaccine development, side affects and storage conditions through accessing those vaccine relative papers.

▶ Scientists: Develop more effective vaccines

                   Study the vaccine side affects and allergic reactions

                   Find appropriate environment to safely store vaccine

▶ Doctors: Develop more effective therapy for patients

                   Cooperate with scientists and provide real life experience

▶ Nurses & Healthcare professionals: Ask medication allergy before covid vaccine
                                        Inform the side effects of covid vaccine

▶ Industry: Restrict the store hours and number of customers in store

                   Promote online shopping and pick up service

▶ Governments (could implement following policy):

                   Arrange people to take vaccines based on age, job and risk

                   Encourage people to wear mask and keep social distance

                   Forbidden non essential social gathering