Covid-19 Impact on Online Food Businesses

Background

Our corpus building project aims to address how Covid-19 prompted a change in the mindset of consumers and business owners, resulting in an increase in online purchases of food and groceries and if this shift persisted post pandemic. Also, the analysis focuses to determine the effect of the pandemic on employment in food industry. This is crucial for grasping consumer behaviour shifts during an outbreak that minimizes physical interaction.

Initially, our aim was to determine if there was a surge in e commerce across various industries from different perspectives of consumers, business proprietors, and governmental entities. Based on the feedback we received during our presentation and in mentor meet up, we redefined our perspective to focus on the above aims related to shift in mindset of consumers and business owners towards increase in online purchase of food or groceries.

After analysing several perspectives of the impact due to Covid-19, we collectively decided to build a corpus on this topic to narrow down our focus to substantiate the shift in consumer and business trend due to the pandemic.

Corpus

Our corpus consists of two sub corpora. The target corpus focusses on attitudes and opinions of people leaning towards online purchases of food and groceries during and post pandemic period. On the other hand, the reference corpus contains articles from news journals highlighting people's experiences and preferences on in-person dining and takeaways. It also contains the business announcements or opinions in expanding food outlets/restaurants during pre-covid period to meet the growing demands of the food industry and enriching customer experience. The total corpus consists of 128 files comprising of 136,295 words. The target corpus contains 56 files comprising of 70900 words and the reference corpus contains 72 files amounting to 65395 words.

The target and reference corpus files can be loaded using AntConc or any other relevant textual analysis tool where target is of post covid era and reference being the pre-covid. After generating the wordlists and the keywords, focus on the words related to target corpus which have high positive Keyness score or statistically significant like "delivery", "ghost", "online", "consumers" and "business" to understand the mindset of the consumers and business owners during the post covid period. To identify the context around each key word, use collocates with high stat value and frequency and its corresponding concordances.

To find further context around the research question, you can also observe the following collocates for each of the keywords in the target corpus:

a. Keyword - “delivery”: collocates with a frequency of 5 - boom, apps, platforms, distribution, model.
b. Keyword - “ghost”: collocates with a frequency of 5 - kitchen, virtual, brands, new, plans.
c. Keyword - “online”: collocates with a frequency of 5- ordering, shopping, platforms, shift, grocery, takeaway.
d. Keyword - “consumers”: collocates with a frequency of 5- shift, convenience, ordering, percent, found.
e. Keyword - “business”: collocates with a frequency of 1 but high stat value - small, pandemic, paralysed, shuttering, vulnerable, strengthen.

The target and reference corpus files can be swapped to identify the words from the reference corpus which contain high positive Keyness score or statistical significance such as “experience”, “casual”, “enjoy” in understanding the mindset of the consumers with respect to dining during pre-covid period.

To find further context around the research question, you can also observe the following collocates for each of the keywords in the reference corpus:

a. Keyword - “experience”: collocates with a frequency of 5 - unique, shopping, dining, provide, customer.
b. Keyword - “casual”: collocates with a frequency of 5 - fast, dining, chains, restaurants, service.
c. Keyword - “enjoy”: collocates with a frequency of 5- dining.

To address the research question on whether there was an effect on employment in food industry due to pandemic, we can utilize concordances around word tokens such as “job” and their collocates to comprehend the employment trends during pre-covid and during covid periods.

For Keyness value analysis we can recommend “Log-Likelihood (4-term)” for keyword statistics and “Hardie’s Log Ratio” for Keyness effect size measure. For collocate analysis we recommend to sort keywords either by using frequency or stat as parameters. We can use regex filter to find patterns in text or search by word tokens. These measures can always be fine-tuned to derive different perspectives around the high frequency or statistically significant key words.

Method

We had to collect substantial number of articles to build our corpora. Therefore, we explored a couple of tools such as Webscrapper.io and Factiva. One of the challenges we encountered with using Webscrapper.io was that it required inspecting individual articles that involved customizing the

approach to fit every format of the website. After exploring both tools, we decided to opt for Factiva as it was more efficient to obtain articles from news journals and blogs.

As a third alternative, we thought of extracting content from YouTube video transcripts which had number of podcasts, interviews and discussion videos related to our focus area. These videos are from consumers, business owners and media which express their opinions around the research question. In order to extract content from Youtube videos, we used an open-source "Python YouTube-Transcript-API". We wrote a Python script for which we provided a YouTube video link as an input and the script extracted all the transcripts of speech and generated a text file as an output.

To get the relevant articles of the target corpus during the covid period, we have used key words like "covid", "pandemic", "online", "impact", "job", "restaurants", "closed" and "groceries" and filtered the dates limiting to the covid period. We also narrowed our industries to food and grocery retail industry to capture the opinions of the consumers and business owners during that period.

On the other hand, to build the sub corpus for the reference, we have used key words like "dine", "restaurant", "jobs", "created", "serene", "grocery", "shopping" to filter the articles from Factiva and restricted the time frame to pre-covid era.

We confined our search to aforementioned words as we initially encountered issues in focusing our direction to find the relevant articles.

Once we collected the relevant articles for both the corpora, we used Jupyter notebook taught in our lectures and lab sessions, to extract the content from HTML files and produce text files. To ensure that we meet the word count across the corpora, we built a python script which counts the number of words across each text file in a directory. This script came in handy to check the total word count in the sub corpora.

We encountered an issue in extracting content from Factiva files where Jupyter notebook was unable to extract those from HTML files. To address this, we made necessary changes to python code to seamlessly extract the required file content. The format of the files downloaded were different for both Google Chrome and Microsoft Edge. Google Chrome downloaded HTML files with an additional folder. We made sure that the program accommodated all files downloaded from both browsers. This approach enabled us to extract the text using the program seamlessly across our team.

Once we included the content from YouTube videos, each file was accounting to approximately 15,000 words which skewed the analysis where one opinion dominated other opinions. Therefore, to strike a balance amongst opinions across all articles, we decided to exclude the content from YouTube videos.

Once the above key decisions were taken, we finalized building our corpus.

Limitations

One of the major limitations of analysing these corpora is addressing the research question on whether it had an effect to the employment in food industry was that the key word "job" did not end up in the keyword list at all. Even after finding the relevant articles where "jobs" are mentioned in them, it is surprising that this keyword was not part of the list. We thought we can address this

limitation by doing concordance analysis on the “job” keyword, but we are not sure if the analysis about the effect on employment is fool proof.

Another limitation with this corpus is that we may not arrive at a conclusion whether the trend of online food purchasing persisted post covid. This is because, the key words related to persistence not being populated in the key word list. However, we can observe a fair idea on its persistence through concordances. But we cannot quantify this in a broader sense.

With collocate measure MI and minimum recommended collocate frequency of five, it is challenging to find substantial amount of collocates for key words of reference corpus in order to find context around our research questions. But it is always possible to reduce the collocate frequency to less than five and change the collocate measure accordingly to generate collocates with statistical significance to substantiate our arguments.

Another limitation is that the corpus collected was from opinions of people across various geographies. Since we have not narrowed it down to a specific geography, we may not have the same base of opinions for both the target and reference corpora for comparing pre-covid and post-covid trends of online food purchasing, to accurately address the research questions.

Appendix

Group Members:

1. Omkar Joshi
2. Sridhar Vannada
3. Dilini De Silva
4. Vajiranath Sudusinghe
5. Dharanya Perinbam

We will share our individual contributions as a separate document.