MINING SOCIAL MEDIA TO IDENTIFY THE IMMEDIATE IMPACT OF COVID-19 PANDEMIC ON THE ROMANIAN RETAILERS: EARLY FINDINGS

Adriana Reveiu[1] **and Denis-Cătălin Arghir**[2]
[1) 2)] *The Bucharest University of Economic Studies, Romania*
E-mail: reveiua@ase.ro; E-mail: denis.arghir@ie.ase.ro

Please cite this paper as:

Reveiu, A. and Arghir, D.C., 2020. Mining Social Media to Identify the Immediate Impact of COVID-19 Pandemic on the Romanian Retailers: Early Findings. In: R. Pamfilie, V. Dinu, L. Tăchiciu, D. Pleșea, C. Vasiliu eds. *6th BASIQ International Conference on New Trends in Sustainable Business and Consumption.* Messina, Italy, 4-6 June 2020. Bucharest: ASE, pp. 1225-1232

Abstract

Social media provides comprehensive support, frequently used by customers to share opinion about products and services, shopping experiences and expectations.

This paper proposes a framework for data acquiring and mining to identify immediate impact of Covid-19 pandemic on the main retailers acting in Romania, as it is expressed by customers' posts published on Twitter social network. A comparative analysis with the year before the pandemic outbreak is included, to assess transformations and challenges generated by the pandemic crisis as they are reflected in customers' messages posted on social media.

Keywords

Social media, retailing, text mining, Twitter, machine learning, Covid-19 crisis, sentiment analysis

JEL Classification

L81, C45, C55

Introduction

Covid-19 pandemic has deeply impacted on people's lives and undeniably has disturbed the global economy. While the long-term impact of this crisis is yet to be established, its immediate impact on retailing is significantly (Roggeveen, 2020), and it is discoverable by mining social media.

Merchants of indispensable goods such as food, healthcare products and groceries have experienced increased opportunities, being request to manage a larger volume of orders and to serve consumers at home, withal they have faced new challenges like inventory, supply chain management, delivery (Roggeveen, 2020), but keeping their environment safe. On the other hand, retailers of non-essential goods are experiencing a significant decline in sales and are pushing to identify new solutions to reach and retain customers, just to survive themselves. As more people have migrated online because of the pandemics, the volume of cyber activities has skyrocketed. More than ever before, customers have shared their opinion, thoughts and practices about their online and in-store shopping experiences by social media.

In this context, the aim of this paper is to create an overview of the main topics approached during pandemic crisis, concerning the main retailers and wholesale traders operating in Romania in contrast with the year before pandemic outbreak. This paper proposes a framework for acquiring and mining Twitter messages (tweets) and applies it in the Romanian context, to identify the most frequently topic of interest, their geographical distribution and customers' sentiments as they infer from customers' messages published online.

Literature review

As social media platforms like Twitter, Facebook, Instagram and YouTube provide mechanism for dynamically collecting data on human behaviour and sentiment, they have been proved useful in a variety of economic fields, to study stock market prices (Si et al, 2013), unemployment, economic indicators in real-time (Poza and Monge, 2020), and consumers' profile (Yelowitza and Wilsona, 2015), to discover emerging business ideas (Lee and Shon, 2019), among others.

Along with digital retailing and advertising, sentiment analysis, service performance measurement and outcomes, social media mining is a major research topic in contemporary retail setting (Souiden et al., 2019). Social media mining has been used in retailing, so far, to reveal popularity of a specific product (Anwar Hridoy et al., 2015), to measure the impact of online retailers' engagement on the brand image and service perception (Ibrahim et al., 2017), to analyse the use of Twitter in supply chain contexts (Chae, 2015), to setup sustainable business models (Onete et al., 2013) or as an effective mean to reach customers (Pop et al.,2019). However, to the best of our knowledge no paper has investigated how the challenges generated during a pandemic crisis have impacted retailers and how their actions have been reflected through the lens of customers messages. This paper tries to fill this gap, aiming to discover the impact of pandemic crisis on customers' topics of interest as reflected from Twitter messages about the main retailers operating in Romania.

Twitter is a micro-blogging platform in which users broadcast hundreds of millions of brief messages daily (Worldometers, 2020), on different topics, being accessed daily, both from public and private locations. One major advantage of Twitter is that its messaging service can be programmatically retrieved via Twitter API (Twitter,2020), both real-time and chronological messages, associated multimedia resources and metadata (Reveiu et al., 2008). Even if in Romania, Facebook has the most popular social network, because of its restrictions in accessing posted messaged, Twitter network has been selected for this research study.

Research methodology and framework

For this research purpose top 10 retailers and wholesale traders operating in Romania were selected, based on their reported turnover in 2018, as reflected by Wall-Street (Wall-Street, 2020), namely: Kaufland, Carrefour, Lidl, Profi, Mega Image, Auchan, Metro, Selgros, Penny and Cora.

Twitter API was used for acquiring messages posted on Twitter social network, through Python programming language and Jupyter Notebook. To filtered out the appropriate messages only, some searching criteria were defined: the keywords correspond to the selected aforementioned retailer names, only texts written in Romanian language were selected, and as geolocation constrain only tweets posted from Romania were captured. An overview of data acquisition workflow is depicted in fig. no. 1.

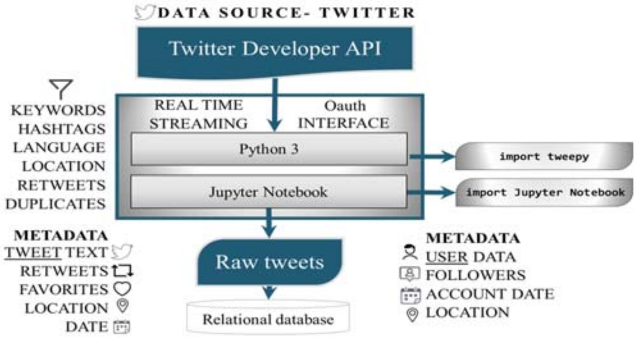


Fig. no. 1 Data acquiring workflow
Source: authors' own design

To identify the immediate impact of Covid-19 pandemics, this analysis considered two time-spans: during-pandemic period which includes the first 40 days of pandemic crisis, between 26th of February (when the first Covid-19 infected person was diagnosed in Romania) and 5 of May 2020, and the pre-pandemic period is the year before pandemic outbreak: 26th of February 2019- 25th of February 2020. The analysis was performed on each data subset allowing to figure out the changes in consumers' concerns in the pandemic time, comparing with the year before (pre-pandemic period).

After completing data collection, 1410 relevant messages were captured for the whole timespan, through Twitter API and stored in the database as raw data. The distribution of the selected tweets on retailers is included in fig. no. 2. The occurrence of retailers in customers' messages posted on Twitter generally followed their market share, as reflected by turnover and presented on the Wall-Street web site.

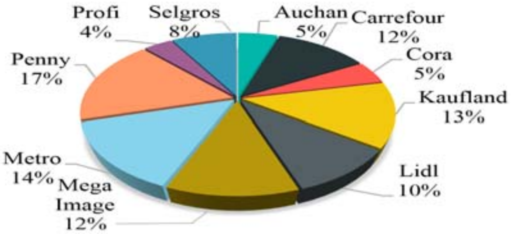


Fig. no. 2 Distribution of collected tweets on retailers
Source: authors' own design based on research results

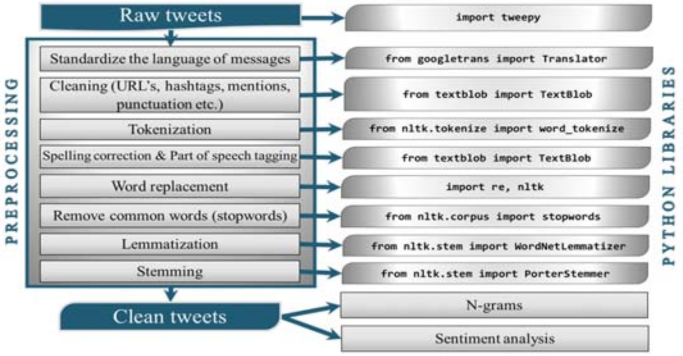
A cleaning procedure was further applied, seeking to remove duplicate messages, emoticons, hashtags, mentions, punctuation signs, non-ASCII characters, stop words and spelling errors. After this stage, 790 tweets, with a relevant content for our analysis, have been finally identified. These were grouped into two data subsets: during-pandemic crisis with 318 messages and pre-pandemics which included 472 messages. Even if the first data subset comprised messages posted during 40 days only and the second one was almost 10 times longer (it lasts for 365 days), the number of collected tweets were nearly the same, proving the increased online activity, on social media, during pandemics.

Text mining analysis framework consists in four main steps: data clearing, natural language processing, sentiment analysis, data clusterization and generates statistically significant topics of interests, as following.

1. In order to mine the text data, a cleaning procedure was firstly accomplished, namely data were tokenized (tweets were divided into word sequences), lemmatized (brought to the

normal dictionary form) and stemmed (word suffixes were removed). To accomplish these operations dedicated Python libraries were used, as briefly depicted in fig. no. 3.

2. Natural language processing libraries were further used to mine the tweets stored in the database, having the purpose to extract the word frequency from Tweeter messages. To accomplish this step word duplicates were identified (by reducing the words to the dictionary form), the parts of the speech were determined (noun, adjective, adverb, verb), special characters were eliminated, typos were corrected, connecting words and irrelevant words were eliminated, according to the flow described in fig. no.3.


Raw tweets
PREPROCESSING
Standardize the language of messages
Cleaning (URL's, hashtags, mentions, punctuation etc.)
Tokenization
Spelling correction & Part of speech tagging
Word replacement
Remove common words (stopwords)
Lemmatization
Stemming
Clean tweets
PYTHON LIBRARIES
N-grams
Sentiment analysis

```
import tweepy
from googletrans import Translator
from textblob import TextBlob
from nltk.tokenize import word_tokenize
from textblob import TextBlob
import re, nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk.stem import PorterStemmer
```

Fig. no. 3 Data cleaning and mining workflow

Source: authors' own design

3. Sentiment analysis uses automated method for evaluating customers' opinion expressed in tweets and for classifying sentiment polarity using three-way classification in: positive, negative and neutral sentiments. Sentiment analysis approached in this paper is based on the Lexicon Sentiment method. To detect sentiment polarity a dictionary of frequently used words was handled. Based on this method the messages including keywords like "good" or "advantageous" are classified as predominantly positive, and if they included words like "bad" or "terrible" the messages are considered predominantly negative.

4. Text based cluster analysis, introduced in fig. no. 4, aims to depict statistically significant topics of discussion related to selected retailers, as they are reflected by Twitter messages. Principal components analysis was firstly used to reduce the dimensionality of our datasets, whilst minimizing loss of statistically significant information. For grouping text messages based on the similarity of their content, the k-means clustering algorithm was selected. The optimal number of required clusters was established using Elbow method. Further, the term frequency is computed, to determine the frequency of word occurrence in messages normalised by their appearance in the entire corpus. Then the values were reduced in a two-dimensional space using principal components analysis.

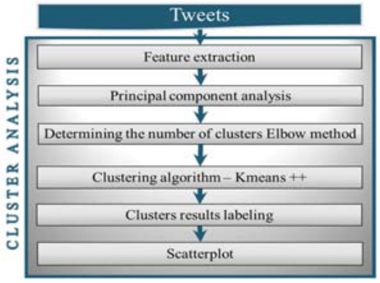

Tweets
CLUSTER ANALYSIS
Feature extraction
Principal component analysis
Determining the number of clusters Elbow method
Clustering algorithm – Kmeans ++
Clusters results labeling
Scatterplot

Fig. no. 4 Cluster analysis workflow

Source: authors' own design

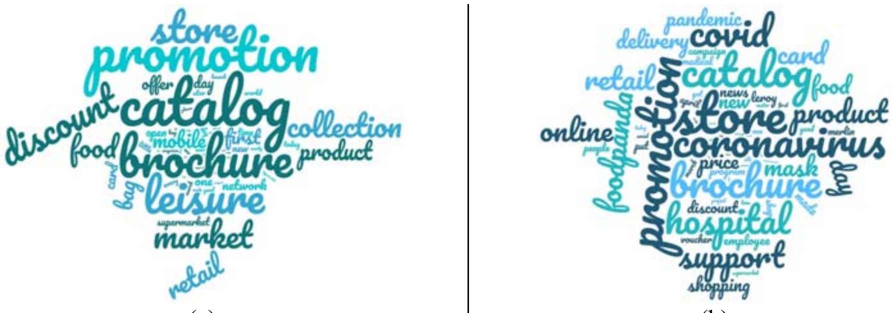
Main research findings

This section presents the main research findings emerged by applying above-mentioned research methodology on both the pre-pandemic and during-pandemic data sets.

During the pandemic crisis, the online discussions related to the main retailers acting in Romania, namely Kaufland, Carrefour, Lidl, Profi, Mega Image, Auchan, Metro, Selgros, Penny and Cora had as authors not only customers, but also retailers, economic magazines and online publications.

In the pre-pandemic period (fig. no. 5a), more specifically the year before (26.02.2019-25.02.2020), messages posted on Twitter mostly reflected customers interest in different areas such as leisure shopping, the discussions being focused on topics related to: discounts, promotions, the use of loyalty card for various bonuses, collections of stickers and surprises for the little ones who come as bonuses for various purchased products or value amounts, and on withdrawing the plastic bags for reasons of pollution and their replacement with biodegradable bags against payment.

Fig. no. 5b presents the most offen approached topics during the first 40 days of pandemic crisis (26.02.2020-5.05.2020), in association with the retailers and wholesale companies. In contrast with the previous timespan, during pandemics the discussions have focused on the online shopping experiences; home delivery of purchased goods; the partnerships between retailers and delivery companies, like FoodPanda; the use of the bank cards to pay for purchases and other utilities through the services provided by retailers, and on donations for supporting hospitals, also.



(a) (b)

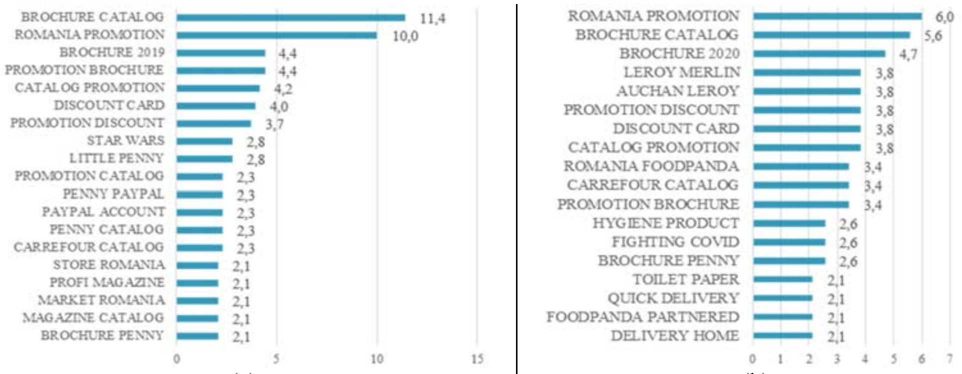
Fig. no. 5 Word clouds of the most relevant terms used by Twitter users in (a) pre-pandemic and (b) during pandemic period

Source: authors' own design based on research results

The interest was obviously focused on the new challenges generated by the Corvid-19 crisis and the protection measures such as medical masks - mandatory for access to some commercial spaces, rubbing alcohol, the construction of the temporary hospital in collaboration between retailer Auchan and Leroy Merlin. The most necessary purchases for the Easter holidays were not neglected, brochures and online catalogues for products promotion also played an important role, bringing attention of those interested by the special shopping program adopted for the state of emergency, Easter and the 1st May. Based on their frequency of occurrence, the fig. no. 5 includes in the world cloud format the most common topics.

To capture the most frequently used combination of words, two bigram charts were designed, as depicted in fig. no. 6, one for each timespan used in our analysis. Based on these charts the interest of customers is more clearly identified, so that before crisis the messages envisaged *promotion brochure*, *catalog promotion, children's collections Star Wars, Kaufland Animatera, Fantasy World*, but also other promotion strategies such as gastronomic live cooking presentations of the chef *Sergiu Nedelea*, or of the *Sitcom Acting - Gourmet*

Improvisational, or personalized offers based on *Discount card*, *Mega Image-Telekom Romania partnership*. During pandemic crisis, the top of identified topics revealed (fig. no 6b) some new combination of words and such some products and services of interest during this period, like: *toilet paper, hygiene product, delivery home, FoodPanda partner, quick delivery, medical equipment, food hygiene, health professional,* and also related to humanitarian actions and foundations, like: *Carrefour foundation, foundation donated, red cross*. The interest remains also towards *promotion brochure*, due to the overlap with the holiday's Easter, but also because the retailers have proposed promotional packages for *fighting COVID*.

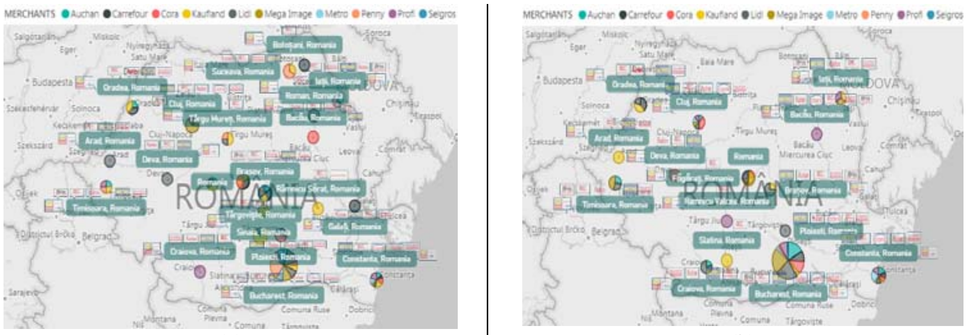


(a) (b)

Fig. no. 6 Bigram for (a) pre-pandemic, and (b) during pandemic period

Source: authors' own design based on research results

From the geospatial perspective, the vast majority of tweets posted during the pre-pandemic period, were identified in the cities where the traders have physical stores. Fig. no. 7 includes geospatial distribution of tweets during the analysed periods.



(a) (b)

Fig. no. 7 Geospatial representation of tweets (a) before and (b) during pandemic period

Source: authors' own design based on research results

While, during pandemic period, the customers' interest on the more convenience stores, like *Mega Image, Profi, Lidl* significantly increased because the long-distance travel was banned and the time slot for reaching the physical markets has become limited. Discussions about large hypermarkets do not disappear at all, but they were more focused on corporate social responsibility - donations to the health system, hospital construction. By contrast, in the pre-pandemic period, the posted tweets discussed about the big players from the market. Following the cluster analysis for all retailers, the messages submitted during pre-pandemic period were clustered into five statically significant groups as they generated the highest level of homogeneity within classes, namely 0.81 from 1. Messages clusterization generates topics of interest. The first cluster includes messages about *Mega-Image - Telekom Romania partnership*, the second cluster groups messages about *customer presence in stores*, the third cluster refers to *collection, promotion and loyalty card*, the fourth group includes messages mentioning *brochures, catalogues and offers* and the largest cluster groups *messages of commercial interest*. Fig. no. 8 includes clusterization results for both periods.

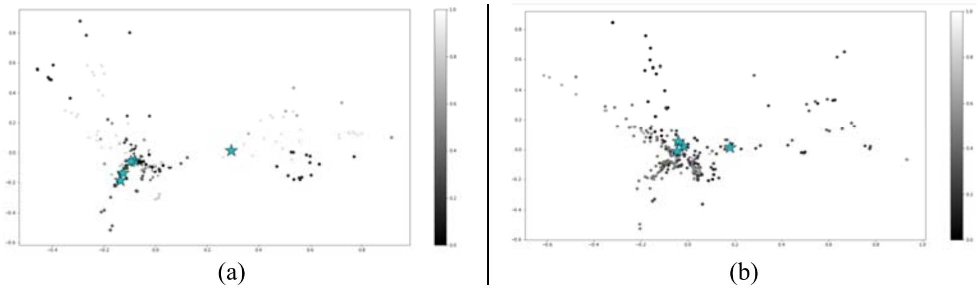

(a) (b)

Fig. no. 8 Cluster analysis (a) before pandemic, and (b) during pandemic crisis
Source: authors' own design based on research results

Even the number of tweets significantly increased during pandemics, they were more homogeneous, from the content perspective. So that, cluster analysis has generated only 4 clusters, at the highest homogeneity level of 0.77 from 1. The content of tweets reflected to a large extent the specificity of this period, as one cluster of messages depict *merchant donations for humanitarian purposes during the pandemic crisis*, the second cluster includes *messages of commercial interest- in connection with pandemic crisis*, the third group includes messages about *brochures, catalogues and offers*, and the smallest one includes messages about *customer presence in stores*

Sentiment analysis revealed a slightly increase, with 3%, of positive perception of all retailers, compared with the previous studied time interval, with a decrease in the same extent of negative perception, as reflected in fig. no.9.

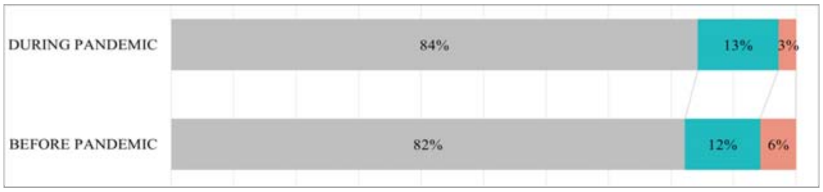


Fig. no. 9 Sentiment Analysis
Source: authors' own design based on research results

Going through the content of the positively classified messages, we noticed that the messages posted both by customers and by merchants urge to responsibility and restraint. Some messages classified as negative, noticed the requirement to wear mask and gloves bought at the entrance of the store at quite high prices, if the customers do not come with them from home. The content of messages classified as neutral announces various promotions, store schedules, online brochures. Examples of messages automatically classified as positive are: "*Carrefour also runs smoothly in Romania. The best prepared hypermarket chains*", "*Mega Image announces salary increases for employees and special bonuses*", "*I want to inform you that I was in the Metro and they took my temperature at the entrance. I'm OK.*"

Conclusions

It is very important for retailer to identify changes in consumers' expectation and habits especially during crisis to be able to anticipate how the landscape will look after the pandemics as these can become the new normal. The results of this research provide evidences that Twitter is a meaningful sensor in detecting patterns in consumers' interests. The framework proposed in this research paper allows to automatize the mining of Twitter messages based on some key terms, the names of retailers, to detect hot topics almost in real time and to mine the text to establish the popularity/opinion/sentiment in different locations from Romania. Even if the amount of acquired tweets was low, because the twitter data was collected following a well-defined procedure this is good enough to demonstrate the utility of proposed framework.

References

Anwar Hridoy, S.A., Ekram, M.T., Islam, M.S., Ahmed, F. and Rahman, R.M., 2015. Localized twitter opinion mining using sentiment analysis. *Decision Analytics*, 2(1), Article number: 8.

Chae, B.K., 2015. Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research. *International Journal of Production Economics*, 165, pp.247-259.

Ibrahim, N.F., Wang, X. and Bourne, H., 2017. Exploring the effect of user engagement in online brand communities: Evidence from Twitter. *Computers in Human Behavior*, 72, pp.321–338.

Imbrea, A., 2019. TOP Cele mai mari retele de retail. Cat vand retailerii internationali in Romania. *Wall-Street*, 19 septembrie 2019, [online] Available at: <https://www.wall-street.ro/special/retailarena/244487/top-care-sunt-cele-mai-mari-retele-de-retail-din-romania.html#gref> [Accessed 30 April 2020].

Lee W.S. and Sohn S.Y., 2019. Discovering emerging business ideas based on crowdfunded software projects. *Decision Support Systems*, 116, pp.102-113.

Onete, C.B., Dina, R. and Vlad, D.E., 2013. Social media in the development of sustainable business. *Amfiteatru Economic*, 15, pp.659-670.

Pop, M.I., Pelău, C. and Stănescu, M., 2019. Reliability of social media platforms and online news as source of information for consumers. In: *5th BASIQ International Conference on New Trends in Sustainable Business and Consumption.* Bari, Italy, 30 May - 1 June 2019. Bucharest: ASE, pp.711-717.

Poza, C. and Monge, M., 2020. A real time leading economic indicator based on text mining for the Spanish economy. Fractional cointegration VAR and Continuous Wavelet Transform analysis. *International Economics*, Article Number: S2110701719302008.

Reveiu, A., Dardala, M. and Smeureanu, I., 2008. A MPEG-21 based architecture for data visualization in multimedia web applications. In: *VIS 2008: International Conference Visualisation*. Proceedings: Visualisation in built and rural environments, pp. 84-89.

Roggeveen, A.L. and Sethuraman, R., 2020. *How the COVID Pandemic May Change the World of Retailing,* Journal of Retailing, [online] Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7183942/> [Accessed 1.05.2020].

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H. and Deng, X., 2013. Exploiting topic based twitter sentiment for stock prediction. In: *51st Annual Meeting of the Association for Computational Linguistic*, pp. 24–29.

Souiden, N., Ladhari, R. and Chiadmi, N.E., 2019. New trends in retailing and services. *Journal of Retailing and Consumer Services*, 50, pp.286-288.

Worldometers, 2020. COVID-19 *Coronavirus Pandemic*, [online] Available at: <https://www.worldometers.info/> [Accessed 30 April 2020].