Social Media Analytics Enabled by Text Mining

Baixin Huang Guowen Song Huanjie Guo Qi Dai Xiaohan Yu Yi Miao

***Abstract*—Social media analytics (SMA) aims to dig out the intentions and sentiment preferences of a particular group of users by collecting data from social media, and use the analytic results to benefit the decision making process of public issues or business functions. Since the breakout of the global pandemic, face mask has become a necessity in people's daily life. In this coursework, we proposes a series of questions to study people's concerning with choosing and wearing face masks. We implement a SMA pipeline, which cleans data collected by Twitter API, and uses name entity recognition and sentiment analysis to parse tweet texts. We successfully answer these questions about face masks, and the final results totally fit the facts.**

***Index Terms*—Social Media Analytics, NLP pipeline, face masks, tweepy**

I. Introduction

Social media analytics (SMA) is being paid more and more attentions since an ever-increasing part of the population start to use social media applications to share their feelings and opinions about the world [1]. Twitter is one of the most popular go-to social media applications to explore what is trending now and a platform for people to make discussions, which makes it a perfect tool to collect data for developing and evaluating informatics [2].

Recently, the UK is gradually resuming people's social interaction, and the outdoor gatherings of either 6 people or 2 households will also be allowed from the Easter holidays. However, many restrictions still remain in place, among which whether to wear face masks is one of the most concerned topics. Therefore, scientifically interpreting public's views on the protective measures during Covid-19, such as wearing masks, is of vital importance for the government to carry out epidemic prevention work.

In this paper, we present three analytical questions to be answered by our social media analytics pipeline: (1) What are people's considerations in the selection of protective products such as appearance, brands and price; (2)What is the usually discount range for masks selling; (3) For each country, what is the proportion of people who tend to think wearing masks is useful to those who think it is not; (4) How do people's attitude towards wearing face masks changes since the outbreak of Covid-19. To answer these questions, we collect a great many relevant tweets with Twitter API to do the sentiment analysis and named entity recognition, which help us learn about people's real thoughts about face masks and other protective measures during Covid-19.

II. Review of related work

A. Named Entity Recognition

The task of named entity recognition (NER) is to identify named entities from unstructured text and classify them into predefined semantic types such as person, location, organization [3]. NER is an efficient standalone tool for information extraction (IE) and it also plays an important role in nature language processing (NLP) field such as question answering, information retrieval and knowledge base construction [4]. The term NER was first defined by Grishman and Sundheim (1996) in the Sixth Message Understanding Conference. Since then, NER has aroused increasing attention (e.g., Tjong Kim Sang and De Meulder, 2003 ; Piskorski et al., 2017).

Considering the techniques applied in NER, there are four main streams [7]. (1) Rule-based approaches, which rely on hand-crafted rules that contains characteristics of the entities as well as terminological information [8]. (2) Unsupervised learning approaches, which rely on unsupervised algorithms using unannotated corpora [9]; (3) Feature-based supervised learning approaches,which rely on supervised learning algorithms n-domain knowledge, gazetteers, orthographic and other features [10]; (4) Deep-learning based approaches,which do not require domain specific resources like lexicons or ontologies and can automatically discover representations needed for the classification [11].

B. Sentiment Analysis

Recently, sentiment analysis is a very popular research field for human-computer interaction practitioners and experts of sociology, marketing and advertising, psychology, economics, and political science. [12] In the past few years, many studies have contributed to the field of sentiment analysis.

A substantial number of sentiment analysis approaches greatly rely on an underlying sentiment lexicon, which is a list of lexical features labeled by their semantic orientation as either positive or negative. One of the most widely used lexicons is LIWC [13], reliable for extracting emotional or sentiment polarity from social media text. But it cannot judge the differences of the sentiment intensity between words. The same drawback also exists in General Inquirer (GI) [14] and Hu-Liu04 opinion lexicon [15].

Not just binary polarity (positive versus negative), many applications developed ways to determine the strength of the sentiment expressed in text, such as SentiWordNet [16] and SenticNet [17]. Besides, typical state of art practices incorporate machine learning approaches to learn the sentiment-relevant features of text.

In this project, we choose VADER (Valence Aware Dictionary and sEntiment Reasoner), a parsimonious rule-based model for sentiment analysis of social media text. [12] It is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon and has been proven to work well on social media style text.

III. Methodology

A. Question 1 2

1) Data Collecting: For this question, we collected target data with keyword "facemasks" from Twitter and filtered those retweet texts. The time span of collected data was from 2021-03-17 to 2021-03-24. Because of the restriction of the $customer\_key$ parameter, we could not download a high volume of data in a single run. Therefore, we repeatedly ran the code block that was responsible for data collection and changed the $cutoff\_date$ and end_date parameter to obtain data from different date. Then, we merged data downloaded in different run together. After that, we manually labelled those texts into five categories and save them in $tweet\_train.csv$.

- Other. This tag was applied to any texts that can not be labelled as the following tags.
- Protective effect. We roughly related any texts containing "covid-19" with this tag because it implied the product was able to protect people from covid-19.
- Appearance. This tag was connected descriptions about the color or the design of face masks.
- Price. This tag was connected with promotions and the price of a product.
- Material. This tag was attached to texts with descriptions regarding the material used by the face mask like "cotton" and "nylon" but it was also attached to texts that described the features of the product like "washable" and "reusable".

2) Data Pre-processing: First, a series of data cleaning strategies were implemented on data read from $tweet\_train.csv$. Then, we split the data into training set and testing set. After this, we generated a word cloud to check whether our data needs any further cleaning. And the result of word cloud was shown in Fig 1.

Fig. 1. Word Cloud

3) Text-Vector Conversion: The third part of task 1 was to convert text to vectors. This contained two subtasks, namely word embedding and sentence representation. The detailed explanations of this part can be found in coursework 1's report.

4) Named Entity Recognition and Classification: In previous labelling process, it would be efficient if we could implement named entity recognition to determine which category a text belonged to. However, the result of NER code (in $NE.csv$) could not support our classification categories due to the lack of domain-specific knowledge. What's more, it was also limited by the characteristics of tweets themselves [18]. For one, there are such a plethora of distinctive named entities types appear in tweets that they will remain low in quantity even though the volume of sample is large. For another, Twitter's 140 character limit makes it difficult to provide sufficient background knowledge for NER. The NER trail for task 1 can be found at "Named Entity Recognition (Part I)". Therefore, we reused the classifier build in coursework 1, the detailed explanations can also be found in coursework 1's report.

For Question 2, we reused the data obtained in Question 1's NER part.And we calculated the proportion of discounts at different level by studying named entities with 'PERCENT' label because we found discounts mentioned in tweets were in the form of percentage.

B. Question 3 4

1) Data Collecting: For Question 3, we need to collect enough tweets from different countries' users, so that we could make more reliable conclusions. However, some countries have fewer Twitter users than other big countries, like the United States and the United Kingdom, and of course they have less discussion on the topic of wearing masks, thus we cannot fetch enough data with tweepy, which only provides tweets of the most recent week. Therefore, to address this problem, we use the same dataset for Question 4, that covers all the tweets with the tag "face masks" from the outbreak last year to today, and select 15 countries in which users give enough opinions on this topic to show the proportion of people in each country who support wearing masks to those who do not.

We use "snscrape", a scraper for social networking services (SNS), to get the collection of historical tweets [19]. And this is also what we want for Question 4 to help us analyze how people's attitudes change since the outbreak of Covid-19.

2) Data Cleaning: In order to clean the data as thoroughly as possible, we take multiple steps on each tweet text we crawled: (1) remove the hyperlinks; (2) remove "@" and the user names mentioned by the tweet; (3) remove "" from the tags; (4) remove escape characters, like "amp;"; (4) remove punctuations and unnecessary line breaks; (5) convert the titles to lowercase; (6) delete the business's advertising tweets; (7) remove duplicate tweets.

3) Topic Modelling: Topic modeling is the process of identifying topics in a set of texts. In this project, we use Latent Dirichlet Allocation (LDA), a form of unsupervised

learning that views text as bags of words, to extract topics from the tweets we collect.

4) Sentiment analysis: We basically use VADER, a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media to finish this phase [12]. Each processed tweet is analyzed by Vader and given a compound score, which is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules and normalized. In this way, we can classify the tweets by the compound scores as either positive, neutral, or negative.

IV. Results

A. Result for Question 1

The accuracy of the classifier on test data set was 0.51. And the proportion of each factors influencing people's options to face masks can be seen in Fig 2. The good news was that around 46.2% people had been aware the horrible consequences of Covid-19 and put protective effect as their priority when choosing face masks. While there remained 36.4% people who failed to realise the importance of protection performance and fixed their eyes on the appearance of face masks. Moreover, about 12.5% people would consider the material of a face mask, leaving only 5% caring about the price of face masks.

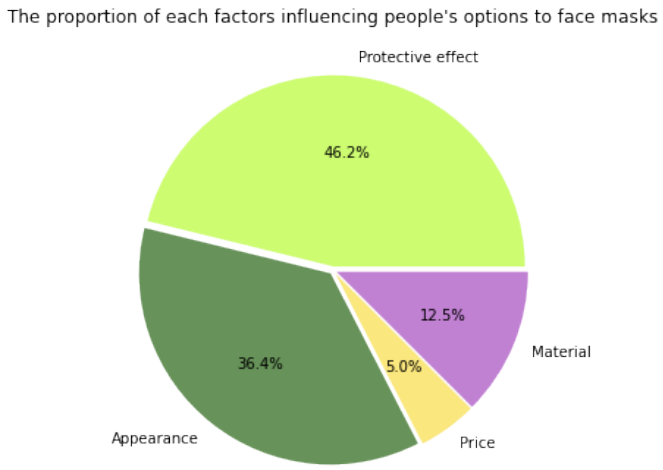

Fig. 2. The proportion of each factors influencing people's options to face masks

B. Result for Question 2

The result of Question 2 was illustrated by Fig 3. As we can see from this bar graph, the median of the discount lied in thirty percent, which was in possession of 25% of total promotion strategies and this is equal to the figure of discount at forty percent. Besides, the discount as fifty percent was also a popular choice with around 20.45% sellers made this decision.

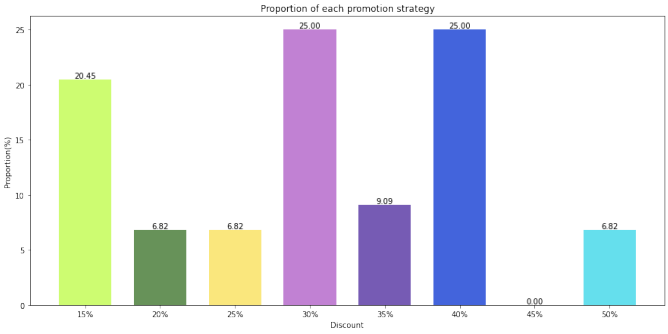

Fig. 3. The proportion of each promotion strategy

TABLE I
A few tweets and its compound score

No.	Text	Compound Score	Sentiment
1	I just want to say to all of you that, I don't want to see lockdown stories on instagram again so please put your mask on.	0.4712	Positive
2	Mask up! The 4th wave is coming.	0.0	Neutral
3	Wearing of face mask has made me stop using lipstick. What has covid-19 protocol changed about you?	-0.296	Negative
4	Face masks could serve another purpose... If you struggle with pollen allergies, wearing a face mask could help.	0.1027	Positive
5	Breathing in mask is easier than breathing on a ventilator.	0.4215	Positive
6	Stop the nonsense coronavirus facemasks!	-0.636	Negative
7	This facemasks mistake is worse than no mask at all!	-0.7901	Negative
8	Wearing a mask is racist and it contributes to climate change.	-0.6124	Negative

C. Topic Visualization

The result of topic modeling is showing below. We choose three topics to build the LDA model, thus they do not overlap and cover most frequent words, which is acceptable in this project. From Fig 4 we can see the most frequent word relate to facemasks is "stay home", which can convey something useful for our analysis.

D. Sentiment Analysis

We randomly select 8 tweets to show the results of sentiment analysis. The table is shown below.

As shown in TABLE 1, the attitudes of most tweets can be correctly recognized. In the texts above, only the No.2, which is a positive sentence, is wrong classified as a neutral one.

E. Result for Question 3

For Question 3, Fig 5 shows the proportions of positive tweets versus negative tweets from 15 countries' users. Among these countries, Japanese Twitter users have the most positive feedback on face masks, and the number of tweets with a supportive attitude is 1.63 times the number of tweets against face masks. And in Ireland, Singapore, Africa, Scotland, Canada and Australia, the ratio exceeds 1, which means people there

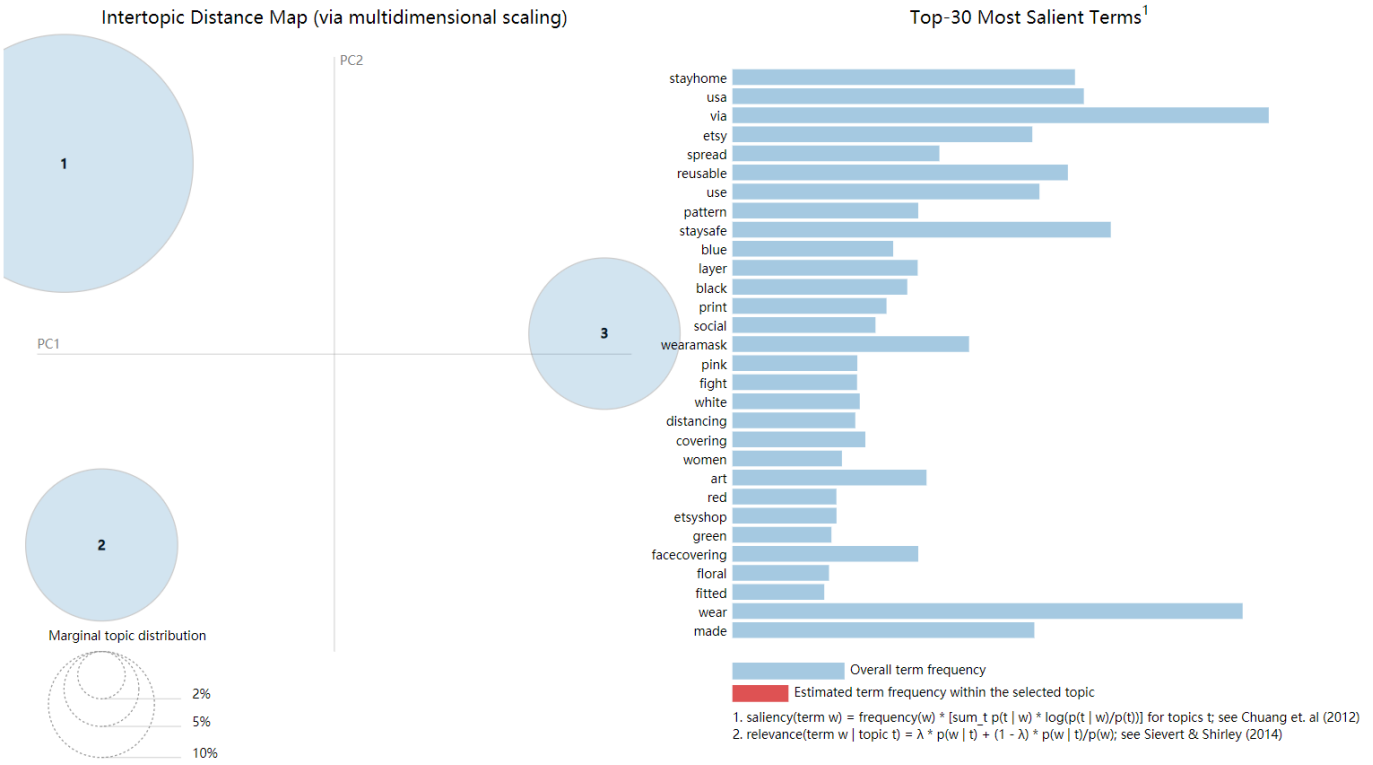

Fig. 4. Topic Modelling

tend to think that wearing face masks is good and harmless. While people from other 8 regions have more negative feelings about masks, especially for Germany, Netherland and the USA, where more people hate wearing masks in the recent year.

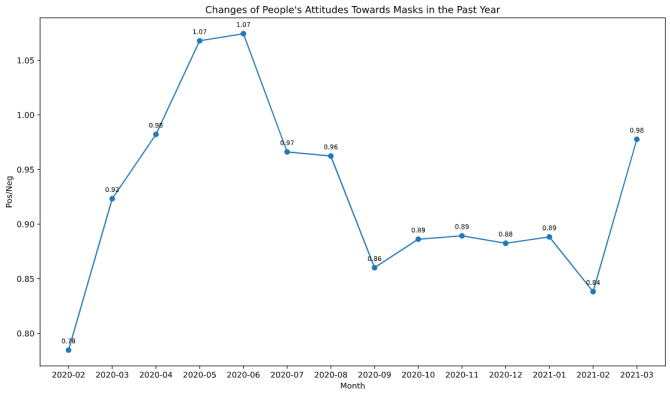

Fig. 5. People's Attitude to Face Masks Changes Since the Outbreak of Covid-19

F. Result for Question 4

The answer to Question 4 is demonstrated in Fig 6, that people's attitude towards face masks have undergone some changes since the outbreak of Covid-19. In February 2020, there is still relatively little discussion about masks, but after the coronavirus has infected a large number of people, the tweets about face masks surged, and those who support wearing masks have expressed a lot of feelings on Twitter. After the summer, the epidemic in China was under control, and people all over the world began to isolate themselves at home for epidemic prevention, and those who oppose or do not appreciate wearing masks conveyed more thoughts on social media. As we can see, recently positive tweets increase again, but still fewer than negative ones.

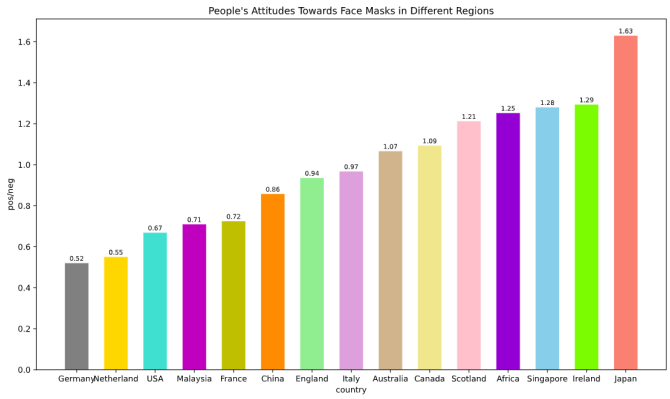

Fig. 6. Proportion of People in Each Country Who Support Wearing Masks to Those Who Do Not

REFERENCES

[1] Stieglitz, Stefan, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. "Social media analytics–Challenges in topic discovery, data collection, and data preparation." International journal of information management 39 (2018): 156-168.

[2] Go, Alec, Lei Huang, and Richa Bhayani. "Twitter sentiment analysis."Entropy 17 (2009): 252.

[3] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification,"Lingvist. Investig., vol. 30, no. 1, pp. 3–26, 2007.

[4] Li, J., Sun, A., Han, J., Li, C., 2020. A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering 1–1.. doi:10.1109/tkde.2020.2981314

[5] Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introductionto the conll-2003 shared task: Language-independent named entity recognition. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, pages 142–147. Association for Computational Linguistics.

[6] Jakub Piskorski, Lidia Pivovarova, Jan˘Snajder, Josef Steinberger, Roman Yangarber, et al. 2017. The first cross-lingual challenge on recognition, normalization and matching of named entities in slavic languages. In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing. Association for Computational Linguistics.

[7] Li, J., Sun, A., Han, J., Li, C., 2020. A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering 1–1.. doi:10.1109/tkde.2020.2981314.

[8] Eftimov, T., Koroušić Seljak, B., Korošec, P., 2017. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. PLOS ONE 12, e0179488.. doi:10.1371/journal.pone.0179488

[9] Qi Peng, Changmeng Zheng, Yi Cai, Tao Wang, Haoran Xie, Qing Li, Unsupervised cross-domain named entity recognition using entity-aware adversarial training, Neural Networks, Volume 138, 2021, Pages 68-77, ISSN 0893-6080,

[10] Ghosh, Souvick. (2016). Feature Based Approach to Named Entity Recognition and Linking for Tweets.

[11] Ronan Collobert, Jason Weston,L eon Bottou, Michael Karlen, Koray Kavukcuoglu,and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. The Journal of Machine Learning Research, 12:2493–2537.

[12] Hutto, Clayton, and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 8. No. 1. 2014.

[13] Tausczik, Yla R., and James W. Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods." Journal of language and social psychology 29.1 (2010): 24-54.

[14] Nielsen, Finn Årup. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs." arXiv preprint arXiv:1103.2903 (2011).

[15] Hu, Xia, et al. "Unsupervised sentiment analysis with emotional signals." Proceedings of the 22nd international conference on World Wide Web. 2013.

[16] Denecke, Kerstin. "Using sentiwordnet for multilingual sentiment analysis." 2008 IEEE 24th international conference on data engineering workshop. IEEE, 2008.

[17] Cambria, Erik, et al. "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives." Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers. 2016.

[18] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11'). Association for Computational Linguistics, USA, 1524–1534.

[19] Get Old Tweets-Python (2020). https://github.com/itsayushisaxena/Get_Old_Tweets-Python.