Protective Headgear: What do people talk about when reviewing products in the context of COVID-19?

Blanka Zana (BZ), *201605645*
Kamilla Natalie Barrett (KB), *202002552*
Blanka Sara Palfi (BP), *201509159*

Supervisors: Rebekah Brita Baglini, Kenneth Enevoldsen
Natural Language Processing for MSc
Aarhus University

Character count: 50,130

Contents

Abstract

This study is set out to explore insights of user experiences indicative of barriers and non-barriers of usage, that can be revealed by modeling topics from product reviews of protective headgears amidst the Covid-19 outbreak in 2020. The data of three different product types of protective headgear (disposable masks, reusable masks and faceshields) is obtained from Amazon (US). The reviews and metadata including star- and usefulness ratings are collected. Topics are modeled using a generative statistical model, SCHOLAR, which can appropriately incorporate the type of metadata collected. Several of the resulting topics provide meaningful insights, which are indicative of certain barriers and non-barriers of wearing protective headgear in the context of Covid-19.

Introduction

(KB)

Since the emergence of SARS-CoV-2 in December 2019, the outbreak has caused great issues for global health, economics, and society (Chakraborty & Maity, 2020). On March 11th, 2020, the World Health Organization (WHO) declared the infectious coronavirus disease a pandemic (World Health Organization, 2020).

To combat the spreading of SARS-CoV-2, a series of countermeasures have been enforced by states and governments worldwide. One of these measures is the introduction of mandatory wear of mask in public places. Masks are proven to be somewhat effective at reducing the travel distance of respiratory droplets caused by coughing and sneezing (Verma et al., 2020).

This paper focuses on gaining information from peoples' opinion that could reveal important aspects of wearing protective headgears, which could potentially be stopping people from wearing them, as often, and efficiently as recommended. It will do so by extracting product reviews, ratings, and usefulness votes of reviews from one of the biggest online stores, Amazon (US), and performing topic modeling on the obtained data.

Firstly, the paper will summarize relevant Covid-19 guidelines for the US public. Secondly, it will introduce the approach and the tool used to explore the data. Thirdly, results are presented and interpreted. Lastly, the findings, limitations, and potential implications for future research are discussed in the context of the paper's scope.

Centers for Disease Control and Prevention (CDC) Guidelines

(BP)

This paper focuses on the American market, for which reason, some relevant Covid-19 guidelines issued by the CDC, a component of the American Department of Health and Human Services are introduced (Centers for Disease Control and Prevention, 2020a).

The CDC recommends using masks covering one's mouth and nose when in public, or around not-household members (Centers for Disease Control and Prevention, 2020b). A private person is urged to use masks made of two or more layers of washable and breathable fabric. Note that surgical masks and N95 masks are not recommended, as they are reserved for medical personnel (Centers for Disease Control and Prevention, 2020c). Within the last few months, an alternative to masks has appeared: faceshields. From November to December 2020 the CDC has updated its recommendation of faceshields from *caution* to *not recommended* (Centers for Disease Control and Prevention, 2020c).

Approach to Research Question

(BZ)

The paper investigates the research question of identifying possible barriers and non-barriers to wearing protective headgear. The study is set out by modelling topics of consumer reviews in English to detect issues around wearing protective headgear experienced by its users. In the analysis three types of popular protective headgears are included: disposable masks, reusable masks, and faceshields. The disposable masks selected are similar to surgical masks, but not approved for medical use at the time, therefore they are in accordance with the CDC recommendation for private people. The selected reusable masks are two ply or more, which is also in compliance with the CDC recommendations. Faceshields are not recommended to be used without a mask, but they are not discouraged to be used with masks.

However, people seem to use faceshields as a substitute for masks. For this reason, it is decided to investigate them alongside masks.

To gauge what people talk about when reviewing protective headgear, topic modelling is conducted on the data. The advantage of using topic modelling to explore the research question, is that the data is readily available and not subject to response biases, such as social desirability bias (Grimm, 2010), which are present in other methods of data collection (e.g., in surveys). Topic modeling is a widely used natural language processing tool used for uncovering latent structures in text corpora. It is often used by people working in the social sciences, for instance detecting opinions of political parties regarding societal issues, identifying biases (Card et al., 2017). A common generative statistical model used in topic modeling is the Latent Dirichlet allocation (LDA; Blei et al., 2003). This method is often used on documents where there is no prior knowledge about a corpus. However, natural corpora can have additional information (*metadata*) stored that can be utilized in topic modeling, e.g., timestamps (Blei and Lafferty, 2006), author information (Rosen-Zvi et al., 2004), or rating (McAuliffe and Blei, 2007). To utilize metadata (MD) present in the dataset used for investigation, a customized LDA model is applied that is appropriate for dealing with such additional information, namely, SCHOLAR, a tool introduced in detail in the following section (Card et al., 2017).

SCHOLAR

(BP)

SCHOLAR (Sparse Contextual Hidden and Observed Language AutoencodeR) is a general neural framework for topic modeling including MD developed by Card et al. (2017). SCHOLAR is available at `https://github.com/dallascard/scholar`. The model is explained below in detail with respect to the traditional LDA (Latent Dirichlet Allocation), prodLDA, SLDA (supervised LDA), SAGE (Sparse Additive Generative Model), and VAE (variational encoder) model that SCHOLAR was built upon.

LDA (Blei et al., 2003) is an unsupervised generative process of inferring topics in a corpus (imagining a generative process in which documents are created based on underlying latent representations of topics). Documents are products of random mixtures over latent topics and each topic is a distribution over all words in the vocabulary of the corpus. In LDA, Dirichlet priors are used on document-word and topic-word distributions. Dirichlet priors are conjugate, meaning that posteriors have the same form as the priors, making updating and tracking of posteriors easy. However, in cases when it would be beneficial to compute the posterior distribution to reparameterize the model (running of alternative models by adjusting parameters using an inference network) Dirichlet priors and posteriors become problematic. This motivated the development of *black box inference methods,* which do not require complex model-specific derivations. Instead, black box inference methods require easy to compute information, which allow for approximate Bayesian inference (Card et al., 2017; Srivastava & Sutton, 2017).

When no MD is added, SCHOLAR is equivalent to prodLDA (Card et al., 2017, p. 4). ProdLDA (Srivastava & Sutton, 2017) aims to tackle the reparameterization challenges associated with inference in topic models described above. The symmetric Dirichlet priors in prodLDA are transformed into logistic normal priors that come handy at reparametrizing (updating of weights and biases in the network with respect to the descriptive mean and variance of the logistic normal distribution). To appropriately account for such differences, in prodLDA, word-topic distributions are unnormalized then transformed with a softmax function (Srivastava & Sutton, 2017). While in LDA each document is a mixture of topics (with *'or'* operation), in prodLDA they are products of topics (with *'and'* operation). This reduces the complexity as a *mixture of topics* accounts for all information available while *product of topics* allows for making a decision based on only a few but relevant dimensions of the problem without covering the full dimensionality (Hinton, 2002).

(KB)

SCHOLAR builds upon two previously developed models when adding MD, SLDA and SAGE (Card et al., 2017). SLDA (Mcauliffe & Blei, 2007) jointly models topics and labels (called responses in the original model in Mcauliffe & Blei, 2007) adding a classification task to topic modeling that functions as supervision since labels are observable data to which the model's predictions can be compared. Conceptually, labels in the SCHOLAR model (also in SLDA) operationalize the research questions, i.e., they define the focus of the topic modeling. As topics and labels are jointly modeled, there is an underlying assumption that topics are informative and representative of labels.

Covariates and interactions can also be modeled in SCHOLAR. This is achieved by incorporating the structure of SAGE appropriate for modeling such MD. Instead of explicit topic-word distributions, SAGE introduces background log-frequencies to model sparse deviations for topic, covariate, and topic-covariate interactions (Eisenstein et al., 2011). This model assumes that covariates, observable MD, have a sparse effect on the relative probability distributions of words with respect to the topics (Card et al., 2017). In comparison to LDA, which eliminates sparsity (zeros when encountering new, unseen words), SAGE induces sparsity in the background log-frequencies and models rare words more accurately with respect to topics, covariates, and their interactions (Eisenstein et al., 2011). Additionally in SCHOLAR, a weight parameter is included on the sparse deviations to account for common words appearing with a similar frequency across documents (Card et al., 2017).

(BZ)

In essence, SCHOLAR can account for two types of MD: labels and covariates. Labels "guide the model to infer topics that are relevant to those labels" (Card et al. 2017, p.6). By adding labels, the model learns unique combinations of topics that can predict labels, thus, it works both as a topic model and a classifier simultaneously (supervised topic model). Covariates on the other hand "induce explicit deviations, leaving the latent variables to account for the rest of the content", i.e., the model attempts to account for meaningful variance in the data with respect to these observable variables (Card et al. 2017, p.6).

As mentioned, SCHOLAR is a neural framework and it operates with VAE (Kingma & Welling, 2014), a sampling-based approximation method for learning and inference (reparameterization). A multilayer perceptron is used to adjust weight and bias terms for the latent topic representations with respect to labels and covariates when included. A reparameterization trick is applied to avoid difficulties with manipulating the posteriors directly. A sampling process from the resulting posterior distributions is utilized, and these are reparametrized with respect to a sampling from an independent source of noise (Card et al., 2017, p. 5). This process allows for optimization using stochastic gradient descent.

Methods

Data Collection

(BP)

To extract structured data from amazon.com *Scrapy 2.4.1* is used in Python 3.7, an application framework for crawling websites (Scrapy 2.4.1, n.d.). Source code at the products' review site is inspected to identify the elements to be used in the scrapy parser for information extraction. In the settings script the bot is identified, then a tutorial (Chauhan, 2020) is used as a guide to build the scrapy parsers and run the spiders for each product to be scraped. In this tutorial, a loop is built to append page numbers at the end of the base urls and a list is created for the spider to crawl through. Instead of identifying subsections of each review (e.g., star ratings, titles), whole reviews are extracted and then a loop is built to structure the information into a dictionary as output. An example of a spider is available at the github link of the project (https://github.com/blanaz/scholar_headgear/blob/main/Scraping_example_spider.py).

Data

(KB)

For each product type, three popular products are chosen from Amazon (US) with the most reviews. For all the products, between 800-2,000 reviews were scraped. *Table 1* summarizes details of the products chosen: brand, type, number of reviews scraped, and link to the product on amazon.com. *Pandas* python package is used to read data from each product and merge them into one dataset (Pandas 1.2.0, 2020). All in all, 13,682 reviews are scraped, 5,190 from reusable cloth masks, 4,493 from disposable surgical masks, and 3,999 from faceshields. Data is collected November 25, 2020. The resulting dataset can be found at the project's github page (https://github.com/blanaz/scholar_headgear/blob/main/DATA_RAW.csv).

Brand	Type	Num. of reviews scraped	Link to product
EnerPlex	Cloth mask	2000 (out of 6788)	https://www.amazon.com/EnerPlex-Premium-3-Ply-Reusable-Face/dp/B088C2WD5F/ref=cm_cr_arp_d_product_top?ie=UTF8
Levi	Cloth mask	1489	https://www.amazon.com/Levis-Re-Usable-Bandana-Print-Reversible/dp/B08KXNVK33/ref=cm_cr_arp_d_product_top?ie=UTF8
Hanes	Cloth mask	1701	https://www.amazon.com/Reusable-Cotton-Face-Mask-Pack/dp/B086WC1LDD/ref=cm_cr_arp_d_product_top?ie=UTF8
Hotodeal	Disposable mask	1703	https://www.amazon.com/Hotodeal-Disposable-Breathable-Protection-Lightweight/dp/B086HMXMH9/ref=cm_cr_arp_d_product_top?ie=UTF8
ANNOR	Disposable mask	1944	https://www.amazon.com/Disposable-Face-Masks-Breathable-Comfortable/dp/B084TQKLCC/ref=cm_cr_arp_d_product_top?ie=UTF8
SfAVEreak	Disposable mask	846	https://www.amazon.com/SfAVEreak-Face-Disposable-Pollution-Protection/dp/B08725VTGK/ref=cm_cr_arp_d_product_top?ie=UTF8
OMK	Face shield	1162	https://www.amazon.com/OMK-Pcs-Reusable-Face-Shields/dp/B087GBJPRZ/ref=cm_cr_arp_d_product_top?ie=UTF8
Maxboost	Face shield	1351	https://www.amazon.com/Maxboost-Protective-Face-Shield-Transparent/dp/B089QYYZFY/ref=cm_cr_arp_d_product_top?ie=UTF8
Sunzel	Face shield	1486	https://www.amazon.com/Fulfillment-Sunzel-Shields-Sponges-Protect/dp/B08K2Q9MD3/ref=cm_cr_arp_d_product_top?ie=UTF8

Table 1: Summary of products chosen to collect reviews from: name of the product, product type, number of reviews scraped, and link to the products' site on amazon.com.

Cleaning the Data

(BZ)

The data is preprocessed using Python 3.7. First, the star rating column is cleaned from all alphabetical characters so that the column only contains the ratings. Reviews that are associated with locations other than the United States are excluded (1940 reviews). Reviews that are written before 2020 are also excluded (there are 0 reviews). Using Nakatani Shuyo's *language-detection* library (version from 03/03/2014) (Danilak, n.d.) non-english reviews are excluded (57). Title and text body of reviews are merged together into one text column. This resulted in a dataset with 11,666 documents (reviews).

Next, the column containing usefulness rating (number of helpful votes a review received from other customers) is cleaned of any alphabetical characters. Usefulness and star ratings (1-5) are both transformed into numeric characters. The number of reviews by star ratings is as follows: 1 star: 2067, 2 stars: 846, 3 stars: 1010, 4 stars: 1360, 5 stars: 6383. It is important to note that the number of positive reviews (4-5 stars) is much higher than of negative reviews (1-2-3), which is 66% and 34% of all data, respectively. Usefulness ratings range between 0 and 2,301 (mean = 2.3, STD = 35.5). This variable is transformed into a binary variable, where all values equal to or above one are considered useful (1), and zero usefulness votes are considered non-useful (0). The number of non-useful reviews (9,289) is significantly higher than that of the useful ones (2,377).

Data is split into a train and test set (80%-20%) using *sklearn* (Pedregosa et al., 2011) and *pickle* (Pickle, 2020) to save the resulting two datasets into the required format of the topic model, json. The data cleaning script (cleaning_data.py) and the resulting dataset (cleaned_data.csv) are available at the project's github page `https://github.com/blanaz/scholar_headgear/`.

Preprocessing of Data

(KB)

SCHOLAR has its own preprocessing pipeline, which includes cleaning text from non-alphabetical characters, lowercasing, tokenizing, and stopwords removal. It also has an inbuilt function to create MD (labels and covariates) files from columns needed for the models and visualizations later on. This process creates multiple files that are used for running the model, including a vocabulary, a matrix containing the word counts per document, and others.

MD

MD consists of the scraped and created variables in the dataset, which are: star rating, product type, usefulness, and brand. These will be shortly described in the following section.

Type

When scraping the products, an additional variable was added to the data, called *type*. The type variable contains information regarding the product category: faceshield, reusable mask, or disposable mask.

Brand

Similarly to the *type* variable, a *brand* variable indicating the manufacturer of the product has been added. It has nine levels, based on the nine products included in the analysis.

Star rating

When leaving an online product review, people are asked to express their overall satisfaction with star ratings. On Amazon, the stars range from one to five, one being the worst score, and five being the best. This indication of overall satisfaction is used as a proxy to review sentiment. Star ratings from one to three are considered "critical" reviews by Amazon, hence these star ratings are expected to have more negative sentiment, while star ratings of four and five are expected to have more positive sentiment.

Usefulness

Usefulness of a review was indicated on a binary scale. There is a large difference observed between the amount of useful and non-useful reviews (2,377, and 9,289, respectively), which calls for inferences regarding usefulness to be made with special care. Thus, non-useful reviews are interpreted with caution considering them a pool of reviews that contain both useful and useless information. Hence, non-useful reviews are considered to be unseen by other customers rather than have been seen and decided upon that they are in fact not useful. This will become important when interpreting topic distributions across useful and non-useful reviews.

Procedure of Model Configuration

(BZ)

To explore patterns in the data, multiple SCHOLAR models are run with different parameter combinations of numbers of topics, epochs, folds and seeds (described further down). First, models without MD are run, to explore patterns in the dataset that could inform the variable selection from the MD later on, and to be able compare models with and without any MD. In the topics of the models without (w.o.) MD, it can be observed that multiple topics are strongly associated with either a positive or negative sentiment. By adding star rating as a covariate this sentiment effect can be somewhat somewhat controlled for. Furthermore, some seemingly neutral topics can be associated with both positive and negative sentiment, thus adding star rating as an interaction with topics allows for a systematic exploration of this relationship. Based on the patterns discovered running models w.o. MD, and given our focus of exploration (barriers to wearing different kinds of protective headgear), type is decided to be included as a label in models with MD to jointly model them alongside topics.

Topics from the two best models, one without any MD (100 epochs, random seed 70, 10 dev-folds, and 9 topics), and one with MD (type as label, star rating as covariate, with an interaction effect, 100 epochs, random seed 42, 10 dev-folds and 9 topics), are listed in the results section. Additional tables and figures can be found in the supplementary material, alongside with all the model outputs.

Parameters

Model selection is carried out evaluating perplexity scores and qualitative interpretation of topics.

Topic number

Topic numbers from 5 to 20 are explored. Models with low number of topics (5) result in incoherent topics, which upon qualitative evaluation, appear to be multiple seemingly unrelated topics put into one. Models with a high number of topics (20) have words appearing across multiple topics, resulting in topics that are identical. The most meaningful topics are found using 9 topics, also having the lowest perplexity score.

Epochs

Determining the number of epochs is a tricky question (an epoch being one cycle of model optimization on the training dataset). By choosing a number that is too high, the model is at risk of overfitting to the training data; by choosing a number that is too low, the model is likely to be underfitted to the data. At first, lower numbers (such as 20, 50), then higher (100, 200) are tried. The qualitatively most comprehensible output is achieved with 100 epochs.

Dev folds

The training data is separated into ten folds, in ten rounds; it is trained on nine folds and validated on the held-out test fold.

Random seed

Different random seeds are used to try to find the best results. In total, 10 different inputs to random seeds are tried, varying between 10 and 100.

Model Evaluation

(BP)

Models are evaluated through calculating perplexity, topic coherence and predictive accuracy of labels (in case of the model w. MD) on the test data additional to qualitative judgement of topic evaluation. Perplexity is calculated on the held-out test data, while topic coherence and in case of the model w. MD, predictive accuracy is gained from running the model on the untouched test-data separated before training the model (Card et al., 2017, p. 6).

Perplexity is a measure of the log-likelihood of the held-out test set (Pleple, 2013a). It can offer an insight into how well the model can predict new data with the given parameters (number of topics). Since SCHOLAR works in a variational Bayesian framework, perplexity score is gained from marginal likelihoods of the observed data given the model, the upper-bound based on the ELBO (evidence lower bound) is reported (Card et al., 2017, p. 6). This score, however, is not meaningful on its own, rather, it serves a basis for model selection.

Topic coherence aims to capture human judgement of meaningful topics by calculating the degree of semantic similarities between frequent words within the topics (Pleple, 2013b). In the Scholar pipeline the top ten words are used for each topic and NPMI (normalized point-wise mutual information) is calculated of all word-pair combinations internally (using the test data) (Card et al., 2017, p. 6). In NPMI probabilities for co-occurrence of word-pair combinations are calculated (Pleple, 2013b).

Results

Models

(BP)

Results from the two best models, introduced above (w.o. MD, and with label, covariate, and interaction of covariate with topic), are explained here in detail in relation to one other. The model w.o. MD has a perplexity of 683.3 on the developmental training set, and 691.7
OBS this is on the test set (20% of entire data)
on the ~~developmental~~ test set. Average coherence score over all topics is .214 (range: .124-.309). The model with MD has an accuracy of label-prediction of .80 in the training set, .77 in the dev test set, and .74 in the test set. This suggests that the model generalizes well and is robust enough not to overfit the training data. It has a perplexity of 657.3 on the dev set, and 635.1 on the test set. Topic coherence over all topics is .193 (range: .061-.296) for the second chosen model w. MD.

Topics

(BP)

Topics were qualitatively assessed by the three authors, combining preexisting knowledge on the products, using the list of most and least frequent words in the topics, and manual search of the corpora. The latter has revealed multiple additional smaller discussion topics around certain aspects of wearing masks and shields; these are referred to as 'subtopics'. Subtopics not only help identifying more barriers, but also inform the process of naming the main topics.

Topics from both models are introduced simultaneously, as most of them appear in both. In the next section, the topics will be listed, and described. The coherence scores are reported for each topic. If the topic is found in the model w. MD, the label probabilities are reported. Both the coherence scores and the label probabilities can be found in the topic title.

Topic no and title w.o. MD /coherence score)	***Topic no. and title w. MD (coherence score)***	***Shared most frequent tokens***	***Not shared most frequent tokens by model w.o. MD***	***Not shared most frequent tokens by model w. MD***
0: Wearing Glasses with Masks (.242)	*8: Wearing Glasses with Masks (.238)*	***'nose' 'glasses' 'wire' 'bridge'***	*'fog' 'mouth' 'around'*	*'gap(s)' 'piece'*
1: Production (.146)	*7: Production and Shipping Related to Disposable Masks (.147)*	***'china' 'usa' 'box'***	*'made' 'poor' 'cheap' 'quality' ' waste'*	*'bag' 'broke(n)' 'attached' 'paid'*
2: Positive Aspects of Online Shopping (.142)	*0: Practical Aspects of Online Shopping (.088)*	***'product' 'fast'***	*'good' 'price' 'great' 'value' 'quality' 'delivery'*	*'money' 'service' 'waste' 'describe'*
3: Size and Fit (.309)	*1: Size and Fit (.296)*	***'large' 'medium' 'adult' 'size'***	*'small' 'big' 'old' 'fit'*	*'year' 'woman' 'old'*
4: Ear Loops (.233)	*4: Value for Money (.075)*	***'money' 'break'***	*'broke' 'put' 'straps' 'ear' 'loops' 'easily'*	*'value' 'price' 'works' 'smell(s)' 'great'*
5: Caring for Reusable Masks (.265)	*5: Caring for Reusable Masks (.281)*	***'white' 'dry' 'dye' 'shrink' 'color' 'cotton'***	*'wash(ed)'*	*'dryer'*
6: Positive Aspects of Masks (.124)	*N/A*	***N/A***	*'comfortable' 'breathable' 'soft' 'best' 'wear' 'love' 'tried' 'masks'*	
N/A	*6: Features of Disposable Masks and Faceshields (.061)*	***N/A***		*'quality' 'weight' 'price' 'expected' 'poor' 'good' 'light' 'arrived'*
7: Faceshields (.296)	*2: Faceshields (.282)*	***'foam' 'plastic' 'shield(s)' 'forehead' 'clear'***	*'face' 'lightweight'*	*'cloudy' 'blurry'*
8: Spanish (.166)	*3: Spanish (.265)*	***Disregarded***	*Disregarded*	*Disregarded*

Table 2: Table comparing most frequent tokens per topic according to compared topics of models w.o. and w. MD.

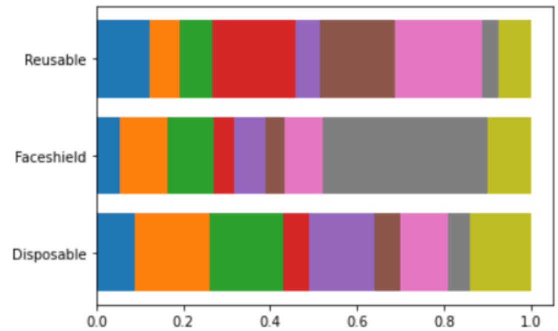

Figure 1a) Distribution of w.o. MD topics by product type.

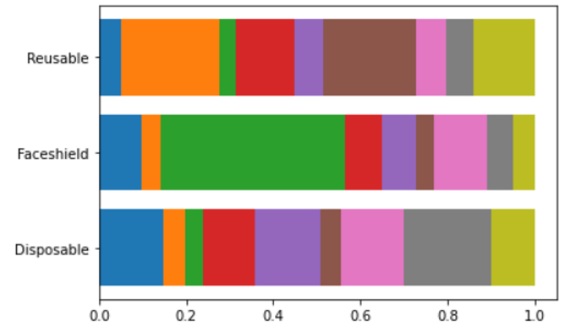

Figure 1b) Distribution of w. MD topics by product type.

Figure 1a) color-code:
Blue: Wearing Glasses with Masks; Orange: Production; Green: Positive Aspects of Online shopping; Red: Size and Fit; Purple: Ear Loops; Brown: Washing of Reusable Masks; Pink: Positive Aspects of Reusable Masks; Gray: Faceshields; Yellow: Spanish

Figure 1b) color-code:
Blue: Practical Aspects of Online Shopping; Orange: Size and Fit; Green: Faceshields; Red: Spanish; Purple: Value for Money; Brown: Caring for Reusable Masks; Pink: Features of Disposable Masks and Faceshields; Gray: Production and Shipping Related to Disposable Masks; Yellow: Wearing Glasses with Masks.

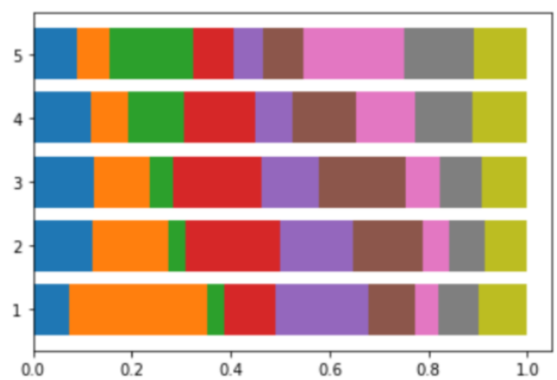

Figure 2a) Distribution of w.o. MD topics by star ratings (1-5). Color-code: See Figure 1a

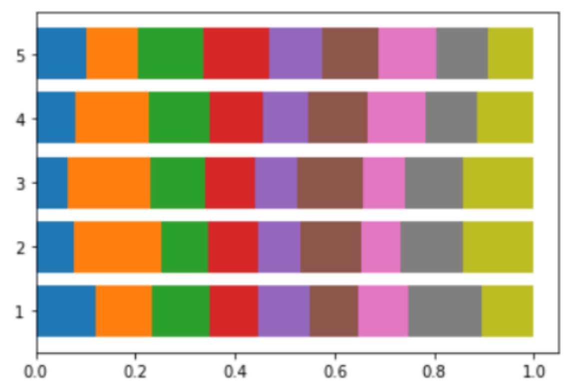

Figure 2b) Distribution of w. MD topics by star ratings (1-5). Color-code: See Figure 1b

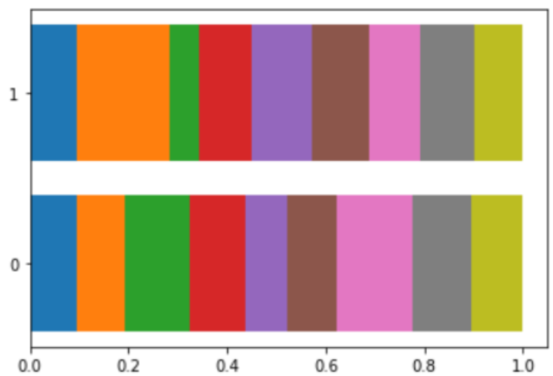

Figure 3a) Distribution of w.o. MD topics by usefulness. 1 = useful. 0 = non-useful. Color-code: See Figure 1a.

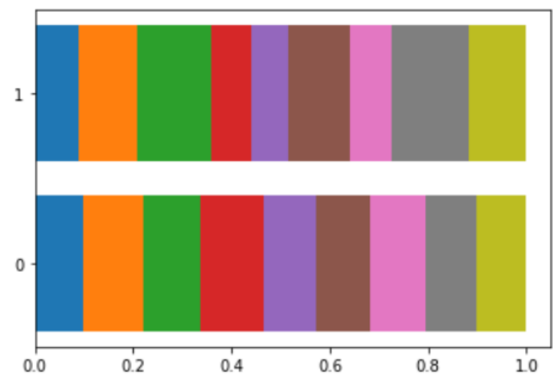

Figure 3b) Distribution of w. MD topics by usefulness. 1 = useful. 0 = non-useful. Color-code: See Figure 1b.

Wearing Glasses with Masks (KB)

Coherence score (CS): w.o. MD: .242 & w. MD: .238

Label probabilities (LP): .65 reusable masks, .33 disposable masks

One topic that repeatedly comes up in various models, as well as in both selected models, is identified as the aspects of *wearing glasses with masks* (*Table 2*). This topic is described with words, such as 'glasses', 'fog', 'wire', 'nose', 'bridge', 'metal'. This is possibly referring to the problem that many people with glasses experience: fogging. The words 'metal', 'bridge', 'nose', and 'wire' refer to the nose wire that certain masks are equipped with; which can prevent the hot air from flowing upward, hence fogging. This topic seems to be unique to masks as the words 'shield' and 'shields' appear on the list of least frequent words of this topic. It can be seen on *Figure 1a,* (blue) and *Figure 1b* (yellow), that this topic is the most predominant for reusable masks and least for faceshields, indicating that the fogging of glasses is a specific aspect of masks. When adding type as a label to be predicted in the second model, this topic is indeed found to be descriptive of masks with .98 probability. Furthermore, the interactions between the star rating and the topics reveal that the words associated with the most negative reviews are 'gap', possibly referring to the gap that allows for the hot airflow that causes fogging; 'useless', possibly referring to dysfunctional nosewires; 'falls' and 'rides' referring to a bad fit of the mask. High star ratings are however associated with words such as 'stays', and 'hooks', possibly referring to a better fit.

Production (BP)

CS: w.o. MD: .146 & w. MD: .147
LP: .97 disposable masks

Words such as 'china','usa', 'box', and 'made' are labeled as the topic of *production and shipping*. To better understand the context of 'china', manual inspection is conducted and a subtopic of *false advertisement* is identified. This subtopic consists of words from a specific discussion around false advertisement and negative reaction to products manufactured in China in the times of the global coronavirus pandemic:

> *I just received these mask and they come from wuhan China ! Very deceiving I bought because it said they were made in the USA! I'm thinking that maybe these were in the USA because they came so fast I paid for two day delivery and in China there's a shortage of face mask ! But I'm still angry that they false advertise! What would you do ? Any advice?*

The topic of *production* is most prevalent in disposable masks (*Figure 1a, orange; Figure 1b, gray*) and in negative, one-star reviews (*Figure 2, orange*). This negative charge can also be observed when looking at the most representative tokens, which seem to be adjectives associated with low quality ('cheap', 'poor', 'waste'). It is worth noting that this topic is almost twice as dominant in reviews deemed useful compared to those that received no votes for usefulness in both models (*Figure 3a, orange; Figure 3b, gray*). The model w. MD shows that this topic is the most descriptive (.93) of disposable masks. When looking at the interactions, low star-rated reviews with this topic talk about 'break', 'snaps', poorly', and 'cheaply'. Interaction with five-star rating is associated with 'hospitals', possibly referring to similarities in product quality with surgical masks worn in hospitals. In five-star reviews, tokens 'everyday', and 'throw' possibly refer to the positive experience of disposable masks used day by day.

> *While these provide great protection, they are also very comfortable. They are very similar to the ones we use in the hospital with COVID patients. Grateful for this level of protection.*

Online Shopping (BZ)

CS: w.o. MD: .142 & w. MD: .088

LP: .73 disposable masks, .23 faceshield

Another topic in both models appears to be focused on aspects of *online shopping*, with words such as ‘price’, ‘product’, ‘fast’, ‘service’ and ‘delivery’ (*Table 2*). Among the least frequent tokens, words that relate to the products themselves (e.g., ‘ear’, ‘loops’, ‘elastic’,’soft’, and ‘comfortable’) are found. The topic itself seems to capture all the surrounding aspects of purchasing (‘good’, ‘price’, ‘product’, ‘great’). There are multiple positive adjectives in the w.o. MD topic. This seems to be the most associated with disposable and the least with reusable masks, however with a slight difference (*Figure 1a, green*), and it is present in all brands (*Figure S1*). This suggests that the topic could be a general aspect of online shopping. The interpretation of sentiment (positive adjectives) is supported by *Figure 2a*, which shows the dominance of this topic in five-star reviews compared to one-star reviews, where it is the least representative of all topics. In the model w. MD, less positive words can be observed, which is expected after adding star ratings as a covariate in the model. The interactions reveal that reviews with low star ratings in this topic are associated with words such as ‘waste’, ‘money’, ‘horrible’, and ‘thin’, all associated with negative descriptions of the products, and not with online shopping. The positive reviews, however, talk about ‘excellent’, ‘fast’, ‘service’, ‘satisfied’, and ‘quick’, all describing the delivery/service. It is important to remember that the number of 4-5 stars reviews is much higher than of 1-2-3-star reviews (66% and 34% of all data, respectively), which could cause that the negative words of some topics are less representative of the topic itself.

Size and Fit of Reusable Masks (KB)

CS: w.o. MD: .309 & w. MD: .296

LP: .97 reusable masks

The *size and fit* detected topic is represented with words such as ‘size’, ‘medium’, ‘small’, and ‘year’ ‘old’ (*Table 2*). This topic is likely more relevant for reusable masks that have different sizes, compared to faceshields and disposable masks which are usually produced in one size. This intuition is supported by the least frequent words, which include 'shield(s)'. A further support for this is the finding that this topic is mostly associated with reusable masks and barely present in the other types (*Figure 1a, red; Figure 1b, orange*). Notably, this topic seems to be the most representative of two- and three-star reviews (*Figure 2a, red; Figure 2b, orange*). An explanation as to why it is less represented in one- and five-star reviews compared to others, could be that a product can have a relatively good or bad fit, but this will not be the reason for the product being perceived as exceptionally good or bad. Rather other topics, such as *Production*, mentioning China (based on the data, people associate China with Covid19) illustrates what makes a product be perceived as exceptionally bad.

Similarly, to the topic of online shopping, this topic is associated with generalities online shopping. However, because a mask's close fit to a person's face, sealing off free flowing air around the mouth, has proven to be essential for the efficiency of a mask, this topic is deemed relevant beyond generalities. A close fit is also recommended by the CDC (Centers for Disease Control and Prevention, 2020c).

The addition of MD to the topic reveals further problematic aspects of sizing, e.g., finding the correct size for one's child. This is observable by tokens 'child', 'age', 'year', 'old', and 'kids' appearing in 1-5-star reviews in the topic-covariate interactions.

> *I have had a hard time finding masks for my 8 year old. Kids's sizes are usually too small, adults too large so I was so happy when my daughter told me these masks fit her perfectly. (...)* (full review, R1 in the Appendix)

Similarly, the word 'beard' appears in the interaction with 5 star reviews. Manual investigation revealed that this is due to customers with large beards having challenges finding fitting masks (mentioned 109 times in the data).

> *These were purchased for my big and tall husband, who has a beard. These are the first masks we have found that fits correctly and offers coverage and comfort. (...)* (full review, R2 in the Appendix).

Ear loops (BP)

CS: w.o. MD: .233

Ear loops of masks are also in focus of the discussions; more specifically, problems associated with poor quality of *ear loops* (‘break’, ‘broke’, ‘straps’). This topic is present only in the model w.o. MD. The model seems to be good at differentiating here between aspects of barriers and non-barriers, as ear loop quality (‘breakage’) is dominant, while other aspects of ear loops e.g., size/fit are not present, making it a very specific topic (*Table 2*). Meanwhile, words associated with function do not appear in the opposite, indicating that breakage of the ear loop actually compromises the function of the masks. Problems with ear loops are mostly associated with disposable masks and least with reusable ones (*Figure 1a, purple*). Further illustrating the problems captured by this topic, that it is the second most prevalent in one-star reviews and least in five-star reviews (*Figure 2a, purple*), while somewhat more represented in useful reviews (*Figure 3a, purple*).

Value for Money (BZ)

CS: w. MD: .075

LP: .77 disposable masks, .12 faceshields, .11 reusable masks

The topic *value for money* is comparable to the topic above, *ear loops*. This topic is also mostly descriptive of disposable masks (.77), and when controlling for sentiment, the two topics share the tokens ‘money’ and ‘break’. *Value for money* is additionally reflected by the tokens ‘value’, ‘price’, and ‘works’. When looking at the interactions, positive sentiment of five-star reviews is reflected with tokens ‘awesome’ and ‘work(s)’, and in one-star reviews, ‘break’, ‘garbage’, ‘waste’, and ‘horrible’ appear. One frequent token of this topic is ‘smell’, which to be made sense of, is manually investigated. The word is mentioned 307 times in the data, of which some may be referring to one’s ability to smell through the mask, while others are referring to the mask itself being of unpleasant odor. Several reviews supporting the latter were found, see an example below.

These masks are horrible and stink bad like something toxic.

Caring for Reusable Masks (KB)

CS: w.o. MD: .265 & w. MD: .281

LP: .99 reusable masks

Another focus of discussions according to both models is the *washing and drying of reusable masks*, with words like ‘cotton’, ‘shrink’, ‘dye’, and ‘dry’ (*Table 2*). The model with MD predicts reusable masks with .99 probability. Tokens reveal a possible barrier in the form of ‘shrink’, ‘dye’ and ‘color’. A manual investigation of those tokens allowed for interpretation of two subtopics: issues or benefits of shrinking masks during washing or drying, and issues relating to the color scheme of masks.

> *They might be a little loose at first, but after a few runs in the washer and dryer they'll snug up just just fine.* (full review, R3 in the Appendix)

The other subtopic is around ‘color’, which reflects that customers do not appreciate receiving products in a different color scheme than ordered, which is generally relevant to shopping.

This is supported by *Figure 1a* (brown), and *Figure 1b* (brown) that shows the topic’s prevalence in reusable as the third most significant topic in this type. From the interactions it can be seen that negative reviews talk about ‘beware’, ‘color’, ‘white’, ‘yellow’, ‘apart’, reflecting the color subtopic, while ‘dryer’,‘dryer’, ‘wash’, and ‘machine’ are possible referring to the fitting of masks by washing/drying.

Positive Aspects of Masks (BP)

CS: w.o. MD: .124

A topic of discussion is associated with *positive aspects of masks* (e.g., ‘best’, ‘love’, ‘soft’, ‘comfortable’, ‘breathable’, *Table 2*). This topic was only found in the model w.o. MD. Some of the least associated words are negative, such as ‘waste’, and ‘poor’, indicating a positive tone for this topic. Other least associated include ‘shield’, and ‘plastic’, which suggest that this topic is descriptive of masks. This topic is most descriptive of reusable masks, where it is the most prominent topic, and least with faceshields (*Figure 1a, pink*). Sentiment is also apparent when plotting the topic distribution across star ratings, where it is the most prevalent topic among five-star reviews, and steadily decreases until one-star reviews (*Figure 2a, pink*).

Faceshields (BZ)

CS: w.o. MD: .296 & w. MD: .282

LP: 0.99 faceshields

Words specific to faceshields (‘foam’, ‘shields’, ‘plastic’, ‘clear’, ‘blurry’, and ‘cloudy’) in general have been found in both models (*Table 2*). There are some positive aspects mentioned, such as ‘clear’ and ‘lightweight’. Intuitively, ‘masks’, ‘size’, and ‘fit’ appear as opposites. It is the most descriptive topic of faceshields (.99) Unsurprisingly, it is the most dominant topic within product category faceshields and is barely represented in the other two categories (*Figure 1a, gray; Figure 1b, green*). It seems to be the most predominant in five-star reviews (*Figure 2a gray; Figure 2b, green*). Investigating the representative tokens of negative reviews, it becomes evident that poor vision through the plastic of a faceshield is a barrier to usage, with tokens ‘cloudy’, ‘blurry’ and ‘scratched’; whereas positive reviews include ‘crystal’.

This topic is found to be somewhat useful by consumers (Figure 3b, green). It is speculated that the novelty and controversy of faceshields make consumers more wary about the purchase of faceshields, for which reason, they could be more likely to rely on others’ experience.

Features of Disposable Masks and Faceshields (KB)

CS: w. MD: .061

LP: .65 disposable masks, .25 masks

This topic only appears in the model with MD. It is largely capturing disposable masks with .65 probability, and partially faceshields with .25 (*Figure 1b, pink*). The topic seems to revolve around the quality of disposable masks. The token ‘quality’ appears over all but one-star rating-topic interaction levels. Positive sentiment in this topic is described with tokens like ‘light’, and ‘good’, ‘price’, and ‘affordable’. Negative sentiment is described with adjectives ‘cheap’, ‘poor’, ‘thin’, and ‘tearing’.

Spanish (BP)

CS: w.o. MD: .166 & w. MD: .265

Even though language detection was run multiple times on the data, it seems like it was unable to detect and delete all Spanish reviews. While this topic is not meaningful for our hypothesis, it is important that our model is able to capture this and group together as a separate topic.

It was decided at the start of the study to only investigate English reviews from the Amazon US website, for which reason all Spanish reviews are excluded from interpretation.

Subtopics found in Alternative Models

(KB)

Some additional subtopics have been identified during the process of model selection, which, while not apparent in the selected models, are included here due to their relevance to the research question (i.e., identifying barriers and non-barriers).

Car

The token 'car' appears in the data 66 times. Manual investigation reveals two subtopics surrounding the token. The first shows that consumers like to keep spare protective headgear in the car. The second subtopic sheds light on an issue of faceshields melting inside cars when temperatures are high. Although there are alternative headgears which do not have this issue, the low melting point of faceshields becomes an inconvenience for consumers who are either momentarily possibly left w.o. protective headgear in the situation. The below review illustrates the issue of faceshields melting:

> *They work and I feel safe shopping with it and a mask. It's so light and clear you forget it's on. But if you leave it in a hot car it'll melt like a Led Zepplin album that you meant to return to the library in 1980.Kinda overpriced for what you get, but whatever.*

Echo

Although this topic is only reflected in four reviews with the token 'echo', it might be relevant for employees of call-service centered jobs. It is in reviews described that an echo occurs when speaking while wearing a faceshield, and that this echo is especially picked up on during digital communication:

> *(…) On the telephone, there's some echo from the plastic shield, so I wouldn't recommend the face shield for anyone doing a lot of phone work. (…) (*full review, R5 in the Appendix*)*

Candle (BP) 1474

The candle blow[1]/lighter test refers to a demonstration of mask function popularized by Bill Nye, a popular media person. In short, if one can blow out a candle while wearing the mask, then the mask is deemed ineffective. The token 'candle' mentioned 47 times, 'flame', mentioned 41 times, and 'lighter test' mentioned 13 times, are all describing this. Example of a review addressing these is below:

> *Its not 3 layer and it failed the blow test with a candle. You can use it but you need 2 mask, to be safe.*

Facial Expression

'Expression' is mentioned in eight reviews, and other tokens like 'be seen' and 'smile' convey the same message: that consumers are happy that they can show their facial expression when wearing faceshields. Although 'expression' or related tokens are not mentioned in masks, it can be inferred that it is a barrier to wearing them, because it is appreciated in faceshields that it allows for communication of facial expressions. Inferring from the reviews, it seems to be most important for consumers who are around kids, teach or having in-person meetings. Please see the reviews below addressing each respectively.

> *I bought this to use as a school teacher. It's a significant improvement over masks, for clarity when I talk, and students can see my facial expressions. (…) (see full review in appendix)*

> *These shields work great for in-person work meetings. You can maintain social distancing and still see each other's facial expressions.*

[1] Read about the candle test here: https://www.npr.org/sections/goatsandsoda/2020/07/24/895090627/coronavirus-faq-what-does-it-mean-if-i-can-blow-out-a-candle-while-wearing-a-mas?t=1607963401286&fbclid=IwAR08KOyNfHQ1A61o12XOi6C0ANXrrRntVOgwK4fwAUFf1A6HmxORqF_FqfA&t=1607964018291

Discussion

Summary of Findings

(BZ)

Overall, the resulting topics provide interesting insights into consumers' experiences with protective headgear in the context of Covid-19. Certain problems of headgears, such as *visibility of faceshields*, *low quality of ear loops on masks*, *fogging of glasses with masks*, *the melting of faceshields*, are detected, some of which with fairly high topic coherence, and straightforward qualitative interpretation. It can be observed that while perplexity scores suggest that the w. MD model is better, coherence scores are higher for the model w.o. MD. The topic with the highest score is of *size and fit*, followed by *faceshields*, and *caring for reusable masks*. The lowest coherence scores are found in topics that had no immediate match in the other model, namely features of *disposable masks and faceshields*, *value for money*, and *positive aspects of masks*. The biggest difference between matched topics of w. and w.o. MD was observed in the pair of *Spanish* topics, and between *ear loops* and *value for mone*y. The other three topics, *wearing masks with glasses, production*, and *online shopping*, all have decent and comparable coherency scores between the topics from the two models, similarly to the topics with higher coherence scores. This suggests that the pairing of topics from the two selected models is reasonable.

The additional subtopics reveal further aspects of headgears, e.g., problems stemming from negative associations with the location of manufacturing (*false advertisement*), specific inconvenience caused by low quality of gears (*car*), a popularly relied on quality-determining factor (*candle test*), and two social factors (*facial expression, and echo*).

The models also provide insights into how people cope with certain problems relating to these headgears, such as shrinking oversized masks, wearing faceshields when facial expressions are of importance, or figuring out an alternative way to assess the efficiency of masks.

(BP)

While getting a general sense of topics in the data already illuminates various aspects, adding *type* as label in the w. MD model helps to answer the research question by allowing for drawing inferences specific to the different product types. Identified possible barriers of reusable masks include *size and fit*, especially when it comes to unconventional mask sizes that fit kids or big beards; the *fogging of glasses*, and *covered facial expressions*. The latter barriers are also present in the case of disposable masks, alongside with *unpleasant odor* of the product, and *unsatisfactory quality of the ear loops*. Barriers of using faceshields are *problems with clarity of vision*, *echoing of speech*, and *low melting point*. Non-barriers include that they allow for *facial expressions* to be seen and are reported by consumers to be comfortable to *wear with glasses*, including not causing fogging of own glasses.

Reflection on Tool Selection

While labels could have been jointly modeled with topics using SLDA, that model does not offer a way for including covariates or interactions, both of which were needed to properly account for the sentiment operationalized as star ratings in the dataset. With SAGE, labels could have been added as covariates, however, that would not allow for modeling them jointly with topics, but as background deviation. SAGE, SLDA, and other alternative models are also appropriate tools for modeling the data. SCHOLAR seems to be the most intuitive choice, given the goals of the study and the diverse MD of the dataset.

Limitations

(*KB*)

At the time of data collection, the CDC guideline for faceshields was to wear with *caution*, as no scientific support for their effectiveness was found. In the meantime, the recommendation has been updated, and currently faceshields are discouraged to be worn by themselves; only recommended to be used in addition to masks. However, including faceshields in the analysis shed light on both barriers- and non-barriers to its own product category and that of masks’. It can be speculated that the appearance of faceshields on the market may be the answer to some of the problems associated with wearing masks.

Another limitation is inherent to the dataset caused by a number of Spanish reviews that failed to be detected during the process of data cleaning. This is at risk for causing noise in the topic modeling, affecting the formation of topics.

Furthermore, better or more valid topics could have been achieved with a bigger and more balanced dataset (in the case of usefulness and star ratings). A more exhaustive search for optimal parameter configurations and MD selections could have been explored, in a more systematic manner, perhaps revealing more robust topics.

The present study contributes to the field of researching the public discourse around wearing protective headgear in two ways: 1) shedding light on some of the barriers associated with wearing different types of headgear, 2) offering some limited insights into the needs of certain populations and specific social situations that could use some attention. When considering specific populations or social situations, explicit take-aways can be formulated based on the insights from this analysis. For example, call centers may benefit from knowing about the echo barrier on two fronts: (i) so that they will not invest in a non-optimal protective headgear for their employees, and (ii) their customers will not experience an annoyance of echo, which could potentially decrease customer satisfaction. It is also noteworthy that specific professions, e.g., teachers, might require alternatives to masks due to their needs of showing and detecting facial expressions. Especially now, when it has been found that faceshields are not safe alternatives.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993–1022.

Blei, D. M., Ng, A. Y., & Jordan, M. I. John D Lafferty. 2006. Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning. ACM (pp. 113-120).

Card, D., Tan, C., & Smith, N. A. (2017). Neural models for documents with metadata. ArXiv Preprint ArXiv:1705.09296.

Centers for Disease Control and Prevention. (2020a, November 5). CDC Organization. Cdc.Gov. https://www.cdc.gov/about/organization/cio.htm

Centers for Disease Control and Prevention. (2020b, November 27). How to Protect Yourself & Others. Cdc.Gov. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html

Centers for Disease Control and Prevention. (2020c, December 21). Your Guide to Masks. Cdc.Gov. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/about-face-coverings.html

Chauhan, B. (2020, June 18). Scraping Amazon Reviews using Scrapy in Python. *Datahut*. https://blog.datahut.co/scraping-amazon-reviews-python-scrapy/

Chakraborty, I., & Maity, P. (2020). COVID-19 outbreak: Migration, effects on society, global environment and prevention. Science of The Total Environment, 728, 138882. https://doi.org/10.1016/j.scitotenv.2020.138882

Danilak, M. M. (n.d.). langdetect: Language detection library ported from Google's language-detection. (1.0.8) [Python; OS Independent]. Retrieved 3 January 2021, from https://github.com/Mimino666/langdetect

Eisenstein, J., Ahmed, A., & Xing, E. P. (2011). Sparse additive generative models of text.

Grimm, P. (2010). Social desirability bias. Wiley international encyclopedia of marketing.

Hinton, G. E. (2002). Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation, 14(8), 1771–1800. https://doi.org/10.1162/089976602760128018

Kingma, D. P., & Welling, M. (2014, April). Stochastic gradient VB and the variational auto-encoder. In Second International Conference on Learning Representations, ICLR (Vol. 19).

Mcauliffe, J., & Blei, D. (2007). Supervised topic models. *Advances in Neural Information Processing Systems*, *20*, 121–128.

Pandas 1.2.0. (2020). *pandas—Python Data Analysis Library*. Retrieved 3 January 2021, from https://pandas.pydata.org/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Pickle. (2020, January). Pickle—Python object serialization—Python 3.9.1 documentation. Docs.Python.Org. https://docs.python.org/3/library/pickle.html

Pleple, Q. (2013a). Perplexity To Evaluate Topic Models. http://qpleple.com/perplexity-to-evaluate-topic-models/

Pleple, Q. (2013b). Topic Coherence To Evaluate Topic Models. http://qpleple.com/topic-coherence-to-evaluate-topic-models/

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. Dans les actes de Proceedings of the 20th conference on Uncertainty in artificial intelligence, 487–494.

Scrapy 2.4.1. (n.d.). *Scrapy 2.4 documentation—Scrapy 2.4.1 documentation*. Retrieved 3 January 2021, from https://docs.scrapy.org/en/latest/index.html

Srivastava, A., & Sutton, C. (2017). Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Verma, S., Dhanak, M., & Frankenfield, J. (2020). Visualizing the effectiveness of face masks in obstructing respiratory jets. Physics of Fluids, 32(6), 061708–7. https://doi.org/10.1063/5.0016018

World Health Organization. (2020, March 11). WHO Director-General's Opening Remarks at the Media Briefing on COVID-19 - 11 March 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

Appendix

Supplementary material

Topic interpretation	*Most frequent words*	*Opposite most frequent words in the word-space*
Topic 0: Wearing glasses with masks	‘nose’ ‘glasses’ ‘wire’ ‘metal’ ‘fog’ ‘mouth’ ‘bridge’ ‘around’	‘shield’ ‘product’ ‘shields’ ‘clear’ ‘quality’ ‘money’ ‘china’ ‘kids’
Topic 1: Production	‘china’ ‘made’ ‘box’ ‘poor’ ‘usa’ ‘cheap’ ‘quality’ ‘waste’	‘comfortable’ ‘soft’ ‘great’ ‘easy’ ‘perfect’ ‘fit’ ‘wear’ ‘fits’
Topic 2: Positive aspects of online shopping	‘good’ ‘price’ ‘product’ ‘great’ ‘value’ ‘quality’ ‘fast’ ‘delivery’	‘ear’ ‘one’ ‘elastic’ ‘loops’ ‘even’ ‘nose’ ‘put’ ‘also’
Topic 3: Size/fit	‘small’ ‘large’ ‘big’ ‘medium’ ‘old’ ‘adult’ ‘size’ ‘fit’	‘price’ ‘product’ ‘covid’ ‘shield’ ‘clear’ ‘protection’ ‘shields’ ‘easy’
Topic 4: Ear loops	‘break’ ‘broke’ ‘put’ ‘money’ ‘straps’ ‘ear’ ‘loops’ ‘easily’	‘fit’ ‘ordered’ ‘received’ ‘large’ ‘small’ ‘soft’ ‘size’ ‘nose’
Topic 5: Taking care of reusable facemasks (washing, drying)	‘white’ ‘dry’ ‘tie’ ‘washed’ ‘cotton’ ‘dye’ ‘wash’ ‘shrink’	‘shield’ ‘covid’ ‘protection’ ‘shields’ ‘clear’ ‘easy’ ‘face’ ‘comfortable’
Topic 6: Positive sentiment towards reusable facemasks	‘comfortable’ ‘breathable’ ‘soft’ ‘best’ ‘wear’ ‘love’ ‘tried’ ‘masks’	‘product’ ‘shield’ ‘elastic’ ‘plastic’ ‘clear’ ‘poor’ ‘waste’ ‘money’
Topic 7: Faceshields	‘shield’ ‘forehead’ ‘foam’ ‘face’ ‘shields’ ‘clear’ ‘lightweight’ ‘plastic’	‘masks’ ‘ear’ ‘nose’ ‘loops’ ‘small’ ‘fabric’ ‘fit’ ‘ears’
Topic 8: Spanish	‘muy’ ‘protection’ ‘bues’ ‘serves’ ‘bueno’ ‘producto’ ‘buenas’ ‘buena’	‘ear’ ‘elastic’ ‘loops’ ‘size’ ‘quality’ ‘large’ ‘big’ ‘put’

Table S1a) w.o. MD topic outputs. The left column is the interpretation of the topics based on the most frequent, i.e. representative words of the topics (middle column) and most frequent words less representative of the topics (on the other end of the spectrum/word-space, right column). shopping, red: topic 3 - size/fit, purple: topic 4 - ear loops, brown: topic 5 - washing cloth masks, pink: topic 6 - cloth masks positive, gray: topic 7 - face shields, yellow: topic 8 - spanish

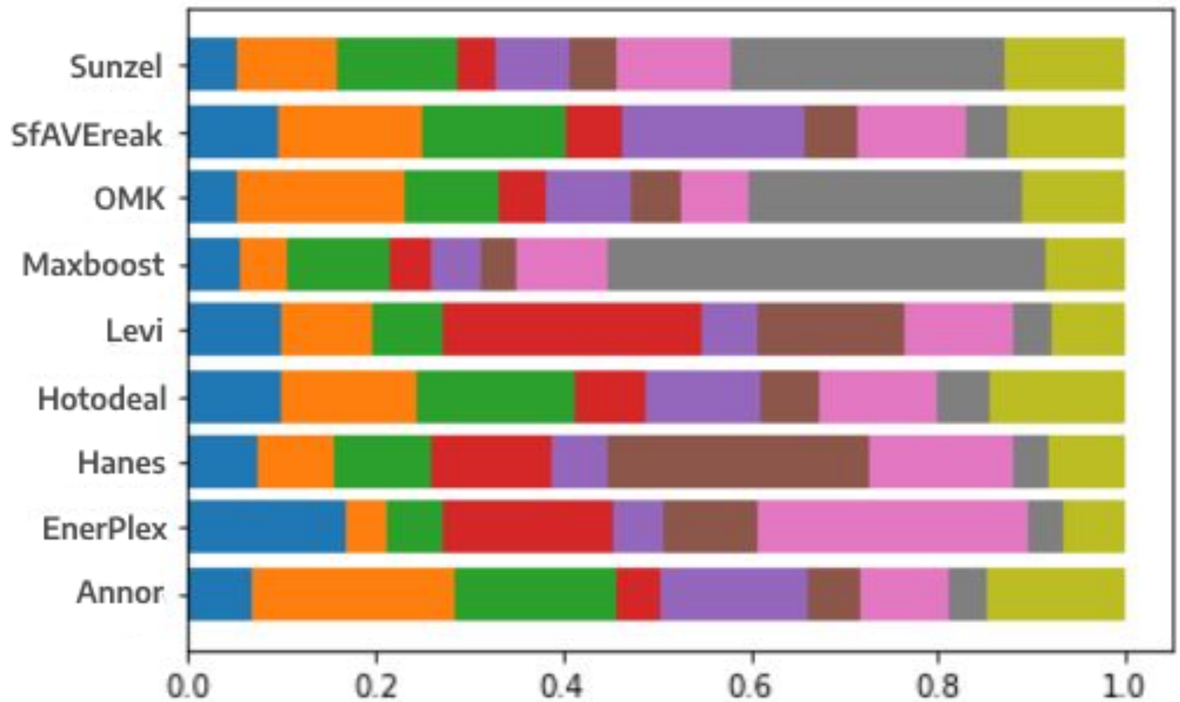

Figure S1a) Distribution of w.o. MDta topics in the chosen products of protective headgears. Color-code: Blue: Wearing Glasses with Masks; Orange: Production; Green: Positive Aspects of Online shopping; Red: Size and Fit; Purple: Ear Loops; Brown: Washing of Reusable Masks; Pink: Positive Aspects of Reusable Masks; Gray: Faceshields; Yellow: Spanish

Topic no.: Interpretation	*Most Representative Tokens*	*Label Probabilities based on Topics in percentage* *1. Disposable Facemasks* *2. Faceshields* *3. Reusable Facemasks*
0: Practical Aspects of Online Shopping	*'product' 'money' 'service' 'waste' 'described' 'excellent' 'cheap' 'fast'*	*(1) .7317* *(2) .2287* *(3) .0396*
1: Size and Fit	*'year' 'adult' 'woman' 'size' 'adults' 'old' 'medium' 'small'*	*(1) .0190* *(2) .0069* *(3) .9740*
2: Faceshields	*'foam' 'plastic' 'shield' 'shields' 'forehead' 'cloudy' 'blurry' 'clear'*	*(1) .0022* *(2) .9968* *(3) .0011*
3: Spanish	*'buen' 'buena' 'muy' 'producto' 'bueno' 'buenas' 'que' 'pedido'*	*(1) .3896* *(2) .0982* *(3) .5122*
4: Value for Money	*'value' 'money' 'price' 'break' 'works' 'smell' 'smells' 'great'*	*(1) .7740* *(2) .1176* *(3) .1084*
5: Caring for Reusable Facemasks	*'white' 'dry' 'dryer' 'dye' 'shrink' 'color' 'tie' 'cotton'*	*(1) .0043* *(2) .0018* *(3) .9939*
6: Features of Disposable Facemasks and Faceshields	*'quality' 'weight' 'price' 'expected' 'poor' 'good' 'light' 'arrived'*	*(1) .6449* *(2) .2532* *(3) .1019*
7: Production and Shipping Related to Disposable Facemasks	*'china' 'usa' 'box' 'bag' 'broken' 'broke' 'attached' 'paid'*	*(1) .9364* *(2) .0191* *(3) .0444*
8: Wearing Glasses with Facemasks	*'wire' 'bridge' 'metal' 'nose' 'glasses' 'gaps' 'gap' 'piece'*	*(1) .3282* *(2) .0216* *(3) .6501*

Table S1b) w. MD topic outputs. The left column is the interpretation of the topics based on the most representative words of the topics (middle column) and the label distribution of the topic (right column).

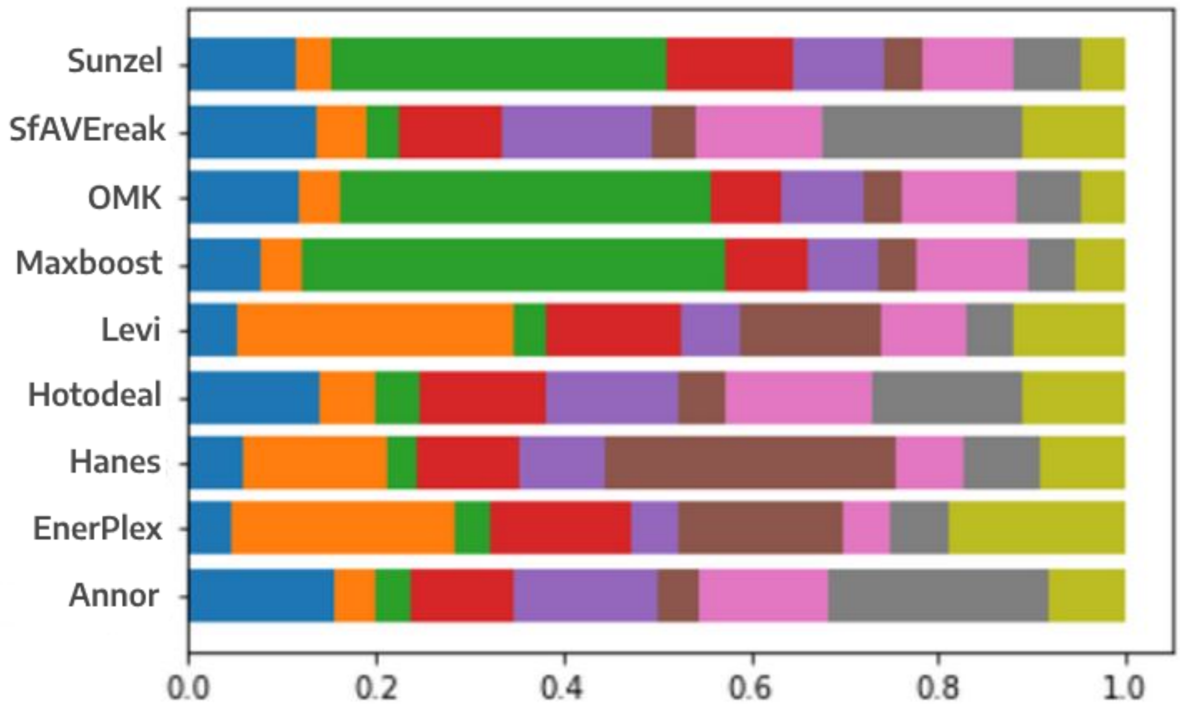

Figure S1b) w. MD topics by brand.

Color-code: Blue: Practical Aspects of Online Shopping; Orange: Size and Fit; Green: Faceshields; Red: Spanish; Purple: Value for Money; Brown: Caring for Reusable Masks; Pink: Features of Disposable Masks and Faceshields; Gray: Production and Shipping Related to Disposable Masks; Yellow: Wearing Glasses with Masks.

Reviews

Review R1:

I have had a hard time finding masks for my 8 year old. Kids's sizes are usually too small, adults too large so I was so happy when my daughter told me these masks fit her perfectly. She can breathe comfortably with them on and is able to go through the school day with no problems. They are reusable and so far, have done well going through the washer and dryer (To make them last longer, I do think I will start handwashing them).Surprisingly, they also fit my 12 year old comfortably so I ended up buying another set of 3.Well worth the money and I may even make a third purchase in the near future!

Review R2:

These were purchased for my big and tall husband, who has a beard. These are the first masks we have found that fits correctly and offers coverage and comfort. We purchased a second set, once we established fit. If you are having difficulty finding a mask to fit, try these! They are great quality and truly xl

Review R3:

They might be a little loose at first, but after a few runs in the washer and dryer they'll snug up just just fine. Bought these after I bought the ones at Costco, which looked cool, but ended up being too small.

Review R4:

These are exactly what I needed. Once you remove the film they are sheer though a bit reflective on the inside. Remember to clean them frequently and they stay that way. The foam gets a bit saturated with sweat during the work day but that is to be expected. They warp when exposed to high heat so don't leave them in the car in summer.

Review R5:

Who could have guessed a year ago that people would be using face shields and masks?I have been using a variety of cloth and surgical masks since the COVID-19 pandemic began, but they get scratchy on the face and uncomfortable around the ears after a full day's use. The full-face shield was an intriguing option, and the MaxBoost shields were a reasonably priced option to try.For quiet office work, they're great - not too much glare, no fogging, comfortable foam padding on the forehead and soft elastic around the head. On the telephone, there's some echo from the plastic shield, so I wouldn't recommend the face shield for anyone doing a lot of phone work.I see why medical professionals like these - they do almost completely protect your face (especially eyes) from droplet sprays. With a mask/respirator, droplet transmission seems very unlikely. I would use these even outside of a pandemic for any chore that involves significant splashing - window washing, car washing, sprinkler maintenance, etc.The PET plastic is not too hot on the face at room temperature - I can see where it would be a problem in hot outdoor situations, though. It is very easy to clean, and resistant to normal household cleaning chemicals like isopropyl alcohol.

Review R6:

I bought this to use as a school teacher. It's a significant improvement over masks, for clarity when I talk, and students can see my facial expressions. It's also nicer than cheaper models, and worth the extra money.

Model Output w.o. MD

9 topics, epochs = 100, random seed = 70

```
python run_scholar.py our_data/ -k 9 --epochs 100 --dev-folds 10 --
seed 70
Loading data
Loaded 9332 documents with 3140 features
Found 9296 non-empty documents
Computing background frequencies
Min/max word counts in training data: 1 3851
Network architecture:
embedding_dim: 300
n_topics: 9
vocab_size: 3140
label_type: None
n_labels: 0
n_prior_covars: 0
n_topic_covars: 0
l1_beta_reg: 0.0
l1_beta_c_reg: 0.0
l1_beta_ci_reg: 0.0
l2_prior_reg: 0.0
classifier_layers: 1
use_interactions: False
Optimizing full model
Epoch: 10 cost= 118.495413114
Epoch: 10; Dev perplexity = 205074.3346
Epoch: 20 cost= 112.805853828
Epoch: 20; Dev perplexity = 3821.7943
Epoch: 30 cost= 113.601724874
Epoch: 30; Dev perplexity = 1466.1061
Epoch: 40 cost= 111.695515134
Epoch: 40; Dev perplexity = 1007.0910
Epoch: 50 cost= 112.268215949
Epoch: 50; Dev perplexity = 863.5041
Epoch: 60 cost= 113.730753099
Epoch: 60; Dev perplexity = 833.9449
Epoch: 70 cost= 113.575112539
Epoch: 70; Dev perplexity = 819.3326
Epoch: 80 cost= 114.257208495
Epoch: 80; Dev perplexity = 803.7956
Epoch: 90 cost= 113.789583053
Epoch: 90; Dev perplexity = 717.3993
Background frequencies of top words:
masks mask face good comfortable great fit quality like product
[0.02611998 0.02391563 0.01784515 0.01662427 0.01660393 0.01529488
```

```
 0.01307017 0.00919728 0.00877675 0.00876318]

Topics:
0: nose glasses wire metal fog mouth bridge around  / shield product
shields clear quality money china kids ; sparsity=0.0000
1: china made box poor usa cheap quality waste  / comfortable soft
great easy perfect fit wear fits ; sparsity=0.0000
2: good price product great value quality fast delivery  / ear one
elastic loops even nose put also ; sparsity=0.0000
3: small large big medium  old adult  size fit  / price covid product
shield clear protection shields easy ; sparsity=0.0000
4: break broke put money straps ear loops easily  / fit ordered
received large small soft size nose ; sparsity=0.0000
5: white dry tie washed cotton dye wash shrink  / shield covid
protection shields clear easy face comfortable ; sparsity=0.0003
6: comfortable breathable soft best wear love tried masks / product
shield elastic plastic clear poor waste money  ; sparsity=0.0000
7: shield forehead foam face  shields  clear lightweight plastic  /
masks ear nose loops small fabric fit ears ; sparsity=0.0000
8: muy protection buen serves bueno producto buenas buena  / ear
elastic size loops quality large big put ; sparsity=0.0000
sparsity in topics = 0.0000
Dev perplexity = 683.2892
Saving document representations
```

Model Output w. MD

Topics = 9, epochs = 100, random seed = 42, Label: type, Covars: star_rating, interaction

```
python run_scholar.py our_data/ -k 9 --epochs 100 --dev-folds 10 --
seed 42 --topic-covars star_rating --labels type --interaction
Loading data
Loaded 9332 documents with 3140 features
Found 9296 non-empty documents
Loading labels from our_data/train.type.csv
Found 3 labels
Loading covariates from our_data/train.star_rating.csv
Train label proportions: [0.37123494 0.22052496 0.4082401 ]
Computing background frequencies
Min/max word counts in training data: 1 3847
Network architecture:
embedding_dim: 300
n_topics: 9
vocab_size: 3140
label_type: None
n_labels: 3
n_prior_covars: 0
n_topic_covars: 5
l1_beta_reg: 0.0
l1_beta_c_reg: 0.0
l1_beta_ci_reg: 0.0
l2_prior_reg: 0.0
classifier_layers: 1
use_interactions: True
Optimizing full model
Epoch: 10 ; cost = 118.892046615 ; training accuracy (noisy) =
0.536273455
Epoch: 10; Dev perplexity = 735213.3268; Dev accuracy = 0.5361
Epoch: 20 ; cost = 112.649073281 ; training accuracy (noisy) =
0.646587785
Epoch: 20; Dev perplexity = 5812.8088; Dev accuracy = 0.6975
Epoch: 30 ; cost = 112.190226339 ; training accuracy (noisy) =
0.670013147
Epoch: 30; Dev perplexity = 1514.0733; Dev accuracy = 0.7589
Epoch: 40 ; cost = 110.220088163 ; training accuracy (noisy) =
0.683638102
Epoch: 40; Dev perplexity = 1084.6346; Dev accuracy = 0.7858
Epoch: 50 ; cost = 109.516050153 ; training accuracy (noisy) =
0.681486793
Epoch: 50; Dev perplexity = 882.8120; Dev accuracy = 0.7847
Epoch: 60 ; cost = 111.779371726 ; training accuracy (noisy) =
0.685072308
Epoch: 60; Dev perplexity = 817.4166; Dev accuracy = 0.7858
```

```
Epoch: 70 ; cost = 111.262131786 ; training accuracy (noisy) =
0.686626031
Epoch: 70; Dev perplexity = 788.4686; Dev accuracy = 0.7783
Epoch: 80 ; cost = 112.753694921 ; training accuracy (noisy) =
0.688896857
Epoch: 80; Dev perplexity = 731.6542; Dev accuracy = 0.7750
Epoch: 90 ; cost = 110.401927993 ; training accuracy (noisy) =
0.686147962
Epoch: 90; Dev perplexity = 671.7626; Dev accuracy = 0.7707
Background frequencies of top words:
masks mask face good comfortable great fit quality product like
[0.02609268 0.0238612  0.01807564 0.01673947 0.0166106  0.01526764
 0.01308364 0.00933964 0.00879025 0.00877669]

Topics:
0: product money described service excellent waste cheap fast  /
light weight glasses wear comfortable breathe soft easy ;
sparsity=0.0000

1: old year adult woman size medium adults small  / break box shield
covid plastic clear piece protection ; sparsity=0.0000
2: foam plastic clear shield shields forehead cloudy blurry  / ear
fabric masks breathable loops disposable straps nose ;
sparsity=0.0000
3: buen buena muy producto bueno buenas que pedido  / ear loops money
elastic price time waste sure ; sparsity=0.0000
4: value smell great price break works money smells  / one ordered
mouth medium tried chin reviews without ; sparsity=0.0000
5: white dryer dye dry shrink color tie cotton  / break shield clear
plastic shields box alternative foam ; sparsity=0.0000
6: quality expected poor arrived good weight light price  / medium
around black make everyone loops chin hurt ; sparsity=0.0000
7: china usa box broken bag broke attached paid  / fits big large
small fit shield soft head ; sparsity=0.0000
8: wire bridge metal nose glasses gaps gap piece  / product price
delivery china clear value came arrived ; sparsity=0.0000
sparsity in topics = 0.0000
Covariate deviations:
1: garbage trash terrible returns junk returned horrible false  /
perfect love excellent highly great glad comfort nice ;
sparsity=0.0000
2: threads tabs nasty unusable shrunk center board disliked  /
excellent highly complaints members nicely works comfy provided ;
sparsity=0.0000
3: downward younger shrink careful okay grader sag hence  / excellent
complaints defective provided enjoy customers save forget ;
sparsity=0.0000
```

```
4: tuck noted dirt wish otherwise fine directed weave  / poor false
garbage elsewhere horrible refund returnable returns ;
sparsity=0.0003
5: love thanks perfect awesome delivery skeptical thank highly  /
poorly false cheaply beware uneven return poor broke ;
sparsity=0.0000
sparsity in covariates = 0.0001
Covariate interactions
(45, 3140)
0:1: waste money junk horrible cheap worst don’t wasted  / excellent
works thank love safety great heavy nice ; sparsity=0.0000
0:2: thin cheap don’t money straps securely comes submit  / excellent
works love described perfect wonderful thank liked ; sparsity=0.0000
0:3: moldable lips goatee worried won’t rigid suggestions lower  /
excellent expectations receive save pleased unable awesome gives ;
sparsity=0.0000
0:4: product awkward effectiveness good points purpose pkg
considerably  / waste junk extremely garbage return plus refund
definitely ; sparsity=0.0000
0:5: excellent fast service product described thank quick satisfied
/ waste horrible trash junk unusable opened never beware ;
sparsity=0.0000
1:1: returnable way child nowhere knock size large individuals  /
love comfy highly breathable several easy pleased boys ;
sparsity=0.0000
1:2: 5’4” marketed forward male rare sewing talked kids  / comfort
highly ripped easy glad couple many ridiculously ; sparsity=0.0000
1:3: huge three definitely runs size cute largest grader  / highly
inside means nope able i’ve clear getting ; sparsity=0.0000
1:4: sizing doubt reviews child imagine older granddaughter fly  /
plus torn never wearing mention returnable shift importantly ;
sparsity=0.0000
1:5: perfectly fits men beard hat medium big year  / returnable
disappointed use useless returns suitable poor throwing ;
sparsity=0.0000
2:1: cloudy plastic fell see blurry scratched flew damaged  / overall
perfect nice bit sturdy added love crystal ; sparsity=0.0000
2:2: medical never severe somewhat envelope sending sure particulate
/ highly pandemic perfect combination clean crystal maxboost nice ;
sparsity=0.0000
2:3: foggy drive velcro nicely serve blurry intended without  /
highly sturdy excellent pandemic used helps love great ;
sparsity=0.0000
2:4: shield protection head slightly giant letters protect band  /
money received maxboost disappointed never scratched fell useless ;
sparsity=0.0000
```

```
2:5: maxboost browser html5 crystal video support shields goggles  /
masks unusable fabric disposable slightly disappointed money worst ;
sparsity=0.0000
3:1: gift metal nunca continues missing pedido itchy rough  / awesome
perfect comfort buen buena thanks reusable thank ; sparsity=0.0000
3:2: floppy levi falls spaces appropriately inside tag sides  / nice
awesome light time perfect pleased buen loves ; sparsity=0.0000
3:3: reads side solid sleeve classic cute tickle bandana  / thanks
buen reusable get loves highly disposable awesome ; sparsity=0.0000
3:4: breathable purpose buen viral served easiest precio depending  /
useless back reversible completely amazon plus apart box ;
sparsity=0.0000
3:5: muy comfortable wear perfect aspects breathable todo sales  /
elastic disappointed returned poor product flimsy side metal ;
sparsity=0.0000
4:1: horrible break smell awful smells terrible garbage waste  /
excellent great value love awesome thank alternative comfy ;
sparsity=0.0000
4:2: ear break smell pay loops strings happening even  / works
excellent glad value awesome lightweight highly pleased ;
sparsity=0.0000
4:3: smell okay better ear elastic stopping kind loop  / excellent
highly family smells everywhere durable awesome share ;
sparsity=0.0000
4:4: otherwise great value moves supply stink money designs  /
ordered waste garbage disappointed horrible also amazon completely ;
sparsity=0.0000
4:5: value works great price easy quantity work awesome  / horrible
one awful waste opened terrible garbage unusable ; sparsity=0.0000
5:1: beware color white disappointed nicely yellow order gonna  /
comfort pleased pocket washable everything members easy snug ;
sparsity=0.0000
5:2: return vary widely rubs can’t became returned wifes  / reusable
comfort highly nicely family pandemic comfy since ; sparsity=0.0000
5:3: synthetic apart staples prepared familys 2nd sitting dryer  /
easy extremely reusable excellent highly find beat supply ;
sparsity=0.0000
5:4: inch length dried vietnam layers gently latex almost  / beware
never returned provided known unusable cheaply crooked ;
sparsity=0.0000
5:5: types dry favorite dryer paint far wash machine  / unusable
beware allergies crooked america visitors product inferior ;
sparsity=0.0000
6:1: poor damaged box cheap quality messed undone inside  / exactly
thank love perfect couple affordable just bit ; sparsity=0.0000
```

```
6:2: poor workout flimsy disappointed exercising get indoor physical
/ liked disposable affordable glad light washes pandemic thank ;
sparsity=0.0000
6:3: brand levi's expected cheap quality tearing seems apart  /
exactly highly excellent promptly thanks right durable perfect ;
sparsity=0.0000
6:4: good price quality timely can't significant beat affordable  /
packed extremely back completely disappointed seller rough scratched
; sparsity=0.0000
6:5: exactly weight good light quality expected price arrived  /
horrible opened able know disappointed levis piece small ;
sparsity=0.0000
7:1: break every breaks strings opened tug card snaps  / best thank
love small layers comfortable pleased found ; sparsity=0.0000
7:2: poorly broke halloween apart attached cheaply already falsely  /
receive pleased soft highly sell although thank n95 ; sparsity=0.0000
7:3: pay silver instantly aluminum know pops products broken  /
highly excellent reusable pleased durable pandemic love defect ;
sparsity=0.0000
7:4: prefer find duty forgetting faster particles virus cheapest  /
attached happened i've disappointed night manufacturer slightest
entire ; sparsity=0.0000
7:5: everyday hospitals stores outstanding purse disposable per throw
/ poorly slightest pops big returns beware larger ripped ;
sparsity=0.0000
8:1: gaps gap money useless worst metal waste floor  / love still
periods helps shield perfect clean comfy ; sparsity=0.0000
8:2: useless ratings falls rides appearance side sucks elsewhere  /
love easy highly purchased needed torn buyer allows ; sparsity=0.0000
8:3: condensation nose conform mostly softest consequently narrow
terms  / heavy waste money product trash pleased false pandemic ;
sparsity=0.0000
8:4: pieces stars piece glasses wear star inside hooks  / waste
poorly disappointed garbage combination breathability complete lips ;
sparsity=0.0000
8:5: chin mouth stays straps hurt talk ears room  / useless worst
waste disappointed clamp elsewhere money spend ; sparsity=0.0000
sparsity in covariate interactions = 0.0000
Dev perplexity = 657.2878
Predicting labels
train accuracy on labels = 0.8018
dev accuracy on labels = 0.7729
Label probabilities based on topics
Labels: Disposable Faceshield Reusable
0: 0.7317 0.2287 0.0396
1: 0.0190 0.0069 0.9740
2: 0.0022 0.9968 0.0011
```

```
3: 0.3896 0.0982 0.5122
4: 0.7740 0.1176 0.1084
5: 0.0043 0.0018 0.9939
6: 0.6449 0.2532 0.1019
7: 0.9364 0.0191 0.0444
8: 0.3282 0.0216 0.6501
Saving document representations

1: garbage trash terrible returns junk returned horrible false ;
sparsity=0.0000
2: threads tabs nasty unusable shrunk center board disliked ;
sparsity=0.0000
3: downward younger shrink careful okay grader sag hence ;
sparsity=0.0000
4: tuck noted dirt wish otherwise fine directed weave ;
sparsity=0.0003
5: love thanks perfect awesome delivery skeptical thank highly ;
sparsity=0.0000
```