Investigating the Use of Topic Modeling for Social Media Market Research: A South African Case Study

Irene Francesca Strydom(✉), Jacomine Grobler, and Euodia Vermeulen

Department of Industrial Engineering, Stellenbosch University, Stellenbosch, Western Cape, South Africa
strydomfrancesca@gmail.com, jacomine.grobler@gmail.com, euodia@sun.ac.za
https://ie.sun.ac.za/

Abstract. Businesses are increasingly investigating the use of data science and machine learning techniques for market research. This paper investigates the use of topic modeling as a tool for social media market research, specifically the influence and impact of this technology within market research practice. As an example of the use of topic modeling, three different topic modeling algorithms are applied to a single dataset extracted from Reddit, and their performance compared. The latent Dirichlet allocation (LDA) algorithm was trained as a baseline and compared to the Correlated topic model (CTM) and the Gibbs sampling for Dirichlet multinomial mixtures (GSDMM) model. The CTM outperformed the LDA model, while the GSDMM was unable to improve on the baseline. The 25 topics produced by the final CTM were investigated in greater detail and interpreted within the context of market research. Although five of these topics did not prove useful, the remaining topics were easily interpreted and divided into six categories related to (1) features, (2) software, (3) acquisition, (4) workouts, (5) physical design, and (6) physiological monitoring. Each category's topics were able to provide valuable insight regarding consumers' opinions about and experiences of the related product.

Keywords: Machine learning · Text mining · Topic modeling · Latent Dirichlet allocation · Correlated topic model · Dirichlet multinomial mixture · Market research · Social media

1 Introduction

Market research is one of the core competencies of most successful businesses today, and the advent of the internet and social media has provided businesses with near limitless sources of data about consumers' habits, opinions and needs. As such, the modern business must leverage this data effectively to differentiate itself and maintain a competitive advantage. Literature reveals that businesses are increasingly investigating the use of data science and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023, LNCS 13957, pp. 305–320, 2023.
https://doi.org/10.1007/978-3-031-36808-0_20

machine learning techniques for market research [9], since traditional market research methods are becoming less effective. These traditional methods suffer from dwindling response rates, are costly and do not necessarily gather input from large audiences [13]. Traditional methods also commonly follow an 'asking' approach, where all initiative is taken by the market researcher. In contrast, social media provides an excellent opportunity for businesses to follow a 'listening' approach. The competitive nature of the booming fitness wearables industry makes a compelling case for businesses in this industry to also listen closely to what their customers are saying about their products on social media. A human movement analysis laboratory has therefore commissioned the research in this paper to determine whether the opinions and experiences of fitness wearable users could reveal potential business opportunities.

An abundance of machine learning techniques are used to analyse the data generated on social media, and these include text mining techniques like topic modeling [1,3]. Topic modeling has mainly been used for information retrieval and organising large corpora of unstructured text. The ease of access to public social media data, however, has led to an increased use of topic modeling to extract themes from social media datasets. The aims of this paper is to assess whether topic modeling is a feasible social media market research tool, what the influence and impact of this technique can be within market research practice, and to conduct a case study comparing the performance of three different topic models as applied to social media data extracted from Reddit.

This paper is significant, since to the best of the authors' knowledge, it is the first to compare the performance of the topic models latent Dirichlet allocation (LDA), the correlated topic model (CTM), and Gibbs sampling for Dirichlet multinomial mixtures (GSDMM) within the context of market research.

The rest of this paper is organised as follows: First an overview of market research and topic modeling is provided, after which data collection and data preparation, are described. Next, the experimental setup and how the algorithms compare are described. The results of the best performing topic modeling algorithm are then interpreted. The paper concludes with findings and recommendations for future work.

2 Relevant Literature

This section presents the literature relevant to the paper. A brief overview of the field of market research and the increased use of data science within this field is presented. The premise of topic modeling, the three topic modeling algorithms relevant to this paper, as well as the chosen method for evaluating the models in this paper are described thereafter.

2.1 Data Science and Market Research

The intersection of data science and market research provide unique opportunities for businesses to innovate and better understand their customers.

The ultimate goal of market research is to enable a business to gain and maintain a competitive advantage through understanding the customers' needs and then catering for those needs. The field of market research is relatively new compared to ancient business fields like transport or logistics, yet it has quickly become indispensable to many corporate businesses around the world. Market research has also changed drastically since its early days, largely owing to the huge amounts of data that researchers now have access to [12]. For the most part, market research relies heavily on quantitative studies, since the data is usually structured and therefore more easily processed and analysed in large volumes [10]. [4], however, explained that customers often have more intricate and nuanced views that cannot be captured by structured data. They use the example of a numerical service rating scale for a hotel's services. If a customer finds room service excellent, but cleaning services terrible, how would they rate their experience? Different customers would give different ratings, but none of these ratings would accurately describe their experience. Bendle and Wang further noted: "By summarizing complex thoughts in a single number, the reasons that informed the choice are lost". A solution might be to add a comment box, in which a customer might express exactly what they feel about each element of their experience. This simple example illustrates the more nuanced nature of qualitative data and how it can provide more insightful and actionable information to a business. The added benefit of conducting research with qualitative *social media* data is that the data is produced without any encouragement from some corporate entity and so provides a unique opportunity to access consumers' genuine and unfiltered opinions. Data mining from social media, therefore, mimics observational studies, which makes it possible to draw out potential associations between the attributes of products and the behaviour of customers. Traditional qualitative studies, however, demand large amounts of time and resources, since nearly every response requires a human to interpret the content [10]. As such, these studies are costly and according to [10], mostly drive quantitative studies or are only used in specific cases. Topic modeling, however, is a method to analyse unstructured, qualitative data on the same scale as structured, quantitative data.

2.2 Topic Modeling

Topic modeling is a text mining technique that is used to identify the latent semantic structure of a corpus, in the form of coherent themes or topics [6]. The assumption is generally that a topic consists of a probability distribution over words or terms, and that each document consists of a probability distribution over topics. Topic models can be classified as unsupervised, supervised, or semi-supervised. This paper focuses solely on three unsupervised topic models. These models are all probabilistic generative models, since they assume that the documents in the training corpus were generated by some probabilistic procedure defined by specific parameters. A topic model aims to infer the set of parameters that best explains the observed terms in documents. Any additional information or assumptions that the analyst has about the corpus is incorporated into the

model through priors, which influence the characteristics of the generative process. Software implementations of topic modeling train the models with a variety of learning algorithms, including collapsed Gibbs samplers and variational inference [2]. These learning algorithms estimate the parameters of the model, since calculating them precisely is intractable.

The topic models considered here are referred to as *bag-of-words* (BOW) models, since only the frequency, and not order, of words in documents are taken into account. In the context of social media, a document could refer to a post, comment or thread. Each document consists of a number of *tokens*, and each token could be assigned a single word, part of a word or multiple words, each referred to as a *term*. The number of unique terms (after pre-processing) is called the vocabulary.

Latent Dirichlet Allocation. LDA, proposed by [6] in 2003, is widely used and considered in literature to be a reliable industry standard for topic modeling. LDA is a probabilistic generative model of a corpus of documents. The assumption is that a document is a multinomial distribution over topics and a topic is a multinomial distribution over terms. In LDA each of these distributions is drawn from a separate Dirichlet distribution $\mathcal{D}$, parameterized by $\boldsymbol{\alpha}$, for drawing the document-topic distributions $\boldsymbol{\theta}$, and $\boldsymbol{\beta}$, for drawing the topic-term distributions $\boldsymbol{\Phi}$. Both α and β are concentration parameters that influence how many topics are present in a document or how many terms make up a topic. The LDA model assumes that the terms in each of the M documents in the corpus are generated by the following process:

1. Sample the term probabilities $\boldsymbol{\Phi}_i \sim \mathcal{D}(\boldsymbol{\beta})$ for each topic i,
2. Sample the topic mixtures of the document $\boldsymbol{\theta}_d \sim \mathcal{D}(\alpha)$ for each document d,
3. Sample the topic assignment $z_{dn} \sim \mathcal{M}(\boldsymbol{\theta}_d)$ for each token n in the document,
4. Sample a term $w_{dn} \sim \mathcal{M}(\boldsymbol{\Phi}_{z_{dn}})$ for each topic assignment z_{dn}.

Correlated Topic Model. CTM is, after LDA, the most widely used topic model and was also proposed by [5] in 2005. Oftentimes the presence of one topic is correlated with the presence of another topic. The underlying Dirichlet distribution used to model the topic mixtures in LDA, however, assumes that the components of these distribution vectors are independent, and therefore fails to directly model the correlations between different topics. To address this problem, CTM utilizes the more flexible logistic normal distribution to model topic proportions for each document. The logistic normal distribution assumes that each document topic distribution $\boldsymbol{\Phi}_i$ is normally distributed with a mean μ_i. Correlations between the different topics are modeled through the covariance matrix $\boldsymbol{\Sigma}$ of the normal distribution. The CTM model assumes a generative process similar to that of LDA, except that the topic proportions for each document are drawn from a logistic normal distribution, rather than a Dirichlet distribution.

Gibbs Sampling for Dirichlet Multinomial Mixtures. Topic models were originally developed for and tested on longer documents such as journal and newspaper articles or Wikipedia pages. The increased interest in analysing social media data has led to the development of new topic models specifically catering to shorter documents, such as those found in social media. [14] conducted a comprehensive survey of short text topic modeling and clustering techniques, each belonging to one of three broad categories, namely: (1) Dirichlet multinomial mixture (DMM) based methods, (2) global term co-occurrence based methods, and (3) self-aggregation based methods. In this paper, the GSDMM algorithm [16] is considered. This algorithm first randomly assigns each document to one of k clusters (hereon referred to as a topic) and then iteratively re-assigns each document in turn with two conditions: (1) the document must be assigned to a topic with more documents than its current topic, and (2) the document must be assigned to a topic where the documents contain similar words to the re-assigned document. If k is large enough, some topics may disappear, while others grow larger. GSDMM can thus infer the optimal number of topics if k is large enough. If k is not large enough, no topics will be discarded. The GSDMM assumes a similar generative process to that of LDA, except that a document is assumed to be generated by a single topic, and not a mixture of topics as assumed by LDA and CTM.

Evaluating Topic Models. Probabilistic models are often evaluated by measuring the log-likelihood of a held-out test [1,7]. The model is trained with the largest part of the data forming the training set, and the rest of the data being held out to test how well the trained model fits this data. For an LDA model the test set is a selection of unseen documents $\boldsymbol{D}$ and the log-likelihood $\mathcal{L}$ represents the probability of the unseen documents being generated given the learned topic and term distributions. A higher log-likelihood therefore implies a better fit. For topic models the perplexity $\mathcal{P}$ of $\boldsymbol{D}$ is more often used as the indicator and is evaluated as:

$$\mathcal{P}(\boldsymbol{D}) = exp(-\frac{\mathcal{L}(\boldsymbol{D})}{f}), \tag{1}$$

where f is the number of tokens in the corpus, and a lower perplexity implies a better model.

The perplexity of a model may be a good indicator of how well a model fits unseen data, but it is often not correlated with how well a human can interpret the topics and their top terms [7,8], since it only measures the probability of observations and not the association of top terms with each other. If human interpretability of topics is important (as would be the case with market research), it is often more useful to determine the *coherence* of topics. The coherence of a topic refers to the semantic similarity of the top terms of the topic and provides a good way to distinguish between topics that are useful for humans and those that are only statistically the best fit for the data.

[8] devised the so called *word intrusion task* to assess the coherence of topics produced by a model through human evaluation. In this task, human subjects are

presented with a list of randomly ordered words, and the subject must identify the word that does not belong with the rest of the words, i.e. the intruder. When the list of words without the intruder fit together, the subjects should easily be able to identify the intruder, implying a topic of good coherence. If none of the words fit together, subjects will most likely choose an intruder randomly, implying a topic of poor coherence. Manually inspecting topics with the word intrusion task can easily become time-consuming, since some large corpora may produce more than 100 topics. Since the main advantage of topic modeling is the automated analysis of large corpora of text, it would not make much sense to again revert back to manual methods to evaluate the topics of a model for market research.

The need to automatically evaluate the coherence of topic models in a way that corresponds with the human interpretability of topics led to the development of so-called *coherence scores* [11]. These scores are calculated by measuring the association of different subsets of the top n words of each topic. Each topic obtains a score, and the overall score is the averaged score of the topic scores. [15] investigated several existing topic coherence measures, and also proposed new measures by combining elements of existing measures. Some of these newly proposed measures - most notably the C_v coherence measure - performed better than existing measures, and are now included in popular topic modeling packages such as *gensim* and *tomotopy*. As such, the C_v coherence score is used to measure the performance of the models in this study.

3 Data Collection

For this paper, data were collected from a forum or community of Reddit users (called a *subreddit*) dedicated to the discussion of a popular brand of wearable fitness devices, hereon referred to as Brand Y. Brand Y's subreddit has more than 49 000 members, and a range of discussions take place daily on this subreddit. Consequently, this subreddit is a treasure trove of information about customers' opinions, experiences and expectations regarding Brand Y's products. A discussion, or thread, typically consist of a single post, which dictates the theme of the discussion, and a number of comments and subcomments made in reply to this post. A single thread could contain anything from a handful to hundreds of comments from any number of users. Posts and comments may contain text, images, videos, and web links. All threads on public subreddits can be accessed via the Reddit application programming interface (API), for which the only requirement is a Reddit account. Several packages in Python, R, and Java have also been developed to interface with the Reddit API, allowing greater ease of access to the data. A combination of *Pushift.io* (an API that interfaces with the Reddit API) and the *praw* package was used to extract the data. The final dataset contains 13 523 posts and 88 927 comments made on Brand Y's subreddit between 1 January 2017 and 1 January 2020.

4 Data Preparation

Before the topic models can be applied to the dataset, the data must first be pre-processed. The objectives of pre-processing are fourfold: (1) to ensure that the model receives input in the correct format, (2) to ensure that the model produces useful and high quality output, (3) to remove irrelevant information in order to speed up the training process, and (4), to address any data quality issues. The effect of each pre-processing step on the corpus and vocabulary size is shown in Table 1. First, each post was combined with all of its corresponding comments to form a thread, and all the threads represent the documents in the corpus. Next, tokenisation was performed, where each document was split into a list of single terms, called tokens. A single list of standard stopwords was then compiled by combining each of the *nltk*, *spacy*, and *gensim* packages' stopword lists. All the terms on this list were removed from the corpus. Even though the standard stopword list contained only 412 terms, it more than halved the size of the corpus (see Table 1). Next, during lemmatisation, each term was reduced to its most basic morphological unit, or lemma. The software used to perform this step uses a parts of speech (POS) tagger to determine what the lemma of a word should be. The word 'connected' for example, is tagged as an adjective and changed to the lemma 'connect', while the word 'connection' is tagged as a noun and the lemma remains 'connection'. Ideally, both words should be changed to the lemma 'connect'. To solve this problem, the POS tagger was disabled. It is noted, however, that a few terms were still not completely reduced to the correct lemma. Thereafter, a list of custom stopwords was compiled and these also removed from the corpus. Custom stopwords include terms that only appear once, terms unique to Reddit (such as 'subreddit' or 'upvote') and terms that carry little semantic meaning in the context, such as 'really' or 'hello'. Bigrams and trigrams were then identified with the *gensim* package and included as single terms in the documents. Examples of identified bigrams and trigrams are: 'battery_life', 'customer_support', and 'health_insurance_discount'. The last step is to convert the corpus into a format that can be fed into the topic model. An id is assigned to each term in the vocabulary and term frequencies in each document are indicated by id-frequency pairs for each document, e.g. the pair (1, 5) indicates that term one appears five times in the document.

5 Modeling

This section describes how the three topic modeling algorithms were applied to the data, and how their performance was evaluated. The LDA algorithm is the most widely used topic modeling algorithm in literature, and as such, was used to establish a baseline against which the CTM and GSDMM algorithms were compared.

Table 1. Pre-processing steps and their effect on the number of terms in the corpus and vocabulary.

Pre-processing Step	Terms in Corpus	Terms in Vocabulary
Combining posts and comments	3 976 985	31 766
Tokenisation	3 976 985	31 766
Standard stopword removal	1 804 221	31422
Lemmatisation	1 804 221	23 962
Custom stopword removal	1 414 956	7 334
Collocation discovery	1 356 137	8 785

The *gensim* package was used to train and optimise the LDA models. *Gensim* includes a feature that allows the LDA training algorithm to learn the optimal values of α and β during training, and this feature was enabled to ensure the best results. Therefore, the only remaining hyperparameter that had to be optimized was k, which was done by conducting a structured statistical test. In the test, a number of models with values of k set to different values between five and 65 were trained. Thirty models were trained for each value of k, and the average of the coherence scores for each set of 30 models was calculated. Each of these models were trained over 100 iterations. Figure 1 shows a plot of the average coherence scores. The average scores confirm that the value of k has a significant effect on the performance measures: $k = 25$ and $k = 30$ have relatively high scores, while there is a decrease in scores for values of k on both sides of the $k = 25$ to 30 interval.

To determine which values for k produce *statistically significant* increases in performance over other values for k, the Mann Whitney U-test, a non-parametric test for independent samples, was used at the 5% significance level. Each sample (set of models with a set value for k) was compared to every other sample and the number of *wins*, *losses* and *draws* for each sample were recorded. A win is recorded if a sample distribution is statistically different from and has a higher mean coherence than another sample, a loss is recorded if a sample is statistically different from and has a lower mean coherence than another sample, and a draw is recorded if there is no statistically significant difference between the samples. The results of each set of Mann Whitney U-tests are shown in Tables 2 and 3, with $k = 30$ yielding the most wins (22), and $k = 25$ in second place with 21 wins.

Table 2. Results of the statistical procedure for $k = 5$ to $k = 20$

k	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Wins	0	1	1	3	3	3	3	7	7	7	7	10	8	10	10	11
Draws	0	1	1	3	3	3	3	5	4	12	15	20	13	20	20	21
Losses	32	30	30	26	26	26	26	20	21	13	10	2	11	2	2	0

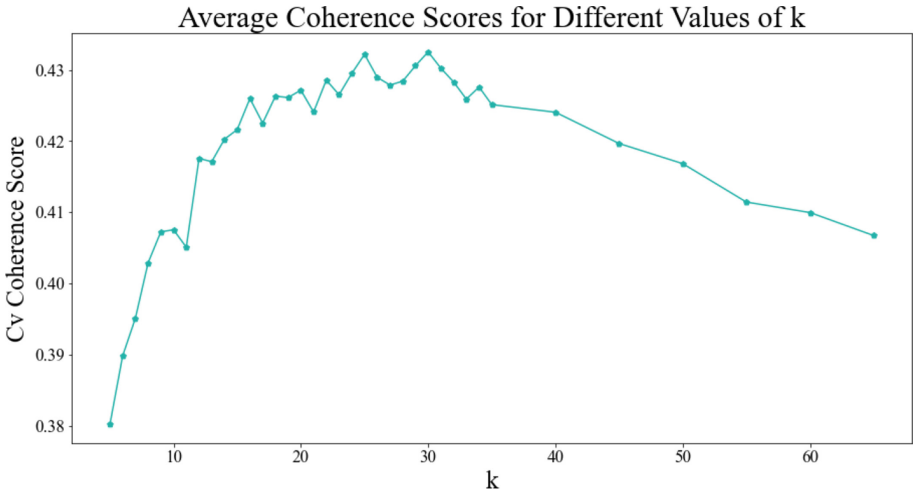

Fig. 1. The average C_v coherence scores for each value of k. There is a steady increase in performance from $k = 5$ to $k = 25$. Two peaks at $k = 25$ and $k = 30$ are observed. After $k = 30$, the performance gradually decreases.

Table 3. Results of the statistical procedure for $k = 21$ to $k = 40$

k	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	40
Wins	9	13	11	15	21	15	14	14	15	22	15	15	10	12	9	9
Draws	16	19	19	17	11	17	18	18	17	10	17	17	20	19	21	14
Losses	7	0	2	0	0	0	0	0	0	0	0	0	2	1	2	9

The results of the statistical testing procedure performed indicated that a model with $k = 30$ yields the best coherence score. It is however noted that $k = 25$ has 21 wins (only one less than $k = 30$), while the four closest runners up ($k = 24, 29, 31$, and 32) have only 15 wins. The sample with $k = 25$ also did not differ from the sample with $k = 30$ in a statistically significant way. Final models for each of the two best performing values of k were therefore trained, and their coherence scores and human topic interpretability compared. The training algorithms for the final models were allowed to run for 1 500 iterations. The final model with $k = 30$ achieved a coherence score of 0.482, while the model with $k = 25$ achieved a score of 0.5. It is interesting to note that even though the statistical procedure indicated $k = 30$ as the optimal choice, a 3.7% higher coherence score is obtained for the final model with $k = 25$. Next, the topics for the two final LDA models were labeled by inspecting the top 20 terms for each topic (i.e. the terms with the highest probabilities of occurring, given a topic) and assigning a label based on the interpreted meaning of the terms. The technical nature of the texts in the corpus and the context in which it was created (social media) mean that domain knowledge is of critical importance. The domain includes not only

wearable fitness devices in general, but also knowledge of the chosen brand's products and services, as well as their features. If the top 20 terms of a topic did not provide any clear, coherent meaning, the topic was marked as a *junk topic*, while if the top 20 terms indicated that more than one concept or meaning is represented by the topic, it is marked as a *mixed topic*. The outcomes of this process are shown in Table 4. The model with $k = 30$ produced worse results than expected. A total of four topics for this model are marked as junk topics, and two are marked as mixed topics. The model with $k = 25$ performed better than expected, with only two topics marked as mixed. The coherence score for the final LDA model with $k = 25$ was the highest, and also produced no junk topics, as opposed to the LDA model with $k = 30$, that produced four junk topics.

Table 4. Outcomes of the topic labeling process.

k	25	30
Labeled	22	24
Mixed	3	2
Junk	0	4

Due to the better interpretability and coherence score of the final LDA model with $k = 25$, the CTM and GSDMM algorithms are also trained with the same value for k. The CTM algorithm is applied to the dataset using the *tomotopy* package. The inputs for the CTM algorithm are k and η. The value for k was set to 25, and the value for η was set to 0.1. A sample of 30 CTM models was trained over 600 iterations each, and an average coherence score of 0.516 was obtained. The final model was trained over 1600 iterations, and achieved a coherence score of 0.527. Finally, a similar procedure is followed to train the GSDMM models using the *gpym_tm* package. The inputs for the GSDMM algorithm are k, α, and β. The value for k was set to 25, while α and β were both set to 0.1. A sample of 30 GSDMM models with $k = 25$ were trained over 25 iterations each, and an average coherence score of 0.404 was achieved. The final model was trained over 50 iterations, and achieved a coherence score of 0.392. The average and final coherence scores for each algorithm are shown in Table 5.

Table 5. Average and final coherence scores for models trained with each algorithm.

Coherence	LDA	CTM	GSDMM
Average over 30 models	0.432	0.516	0.404
Final Model	0.500	0.527	0.392

The Mann-Whitney U-test was then used to compare the results of each of the algorithms and determine which performed best. The results of the procedure are shown in Table 6. From Tables 5 and 6, it is clear that based on C_v coherence scores, the CTM algorithm performed better than the LDA algorithm, while the GSDMM performed worse than the LDA algorithm.

Table 6. Results of the statistical test.

Algorithm	LDA	CTM	GSDMM
Wins	1	2	0
Draws	0	0	0
Losses	1	0	2

6 Interpretation of Topics

The CTM model performed the best, with an average coherence score of 0.516 for the 30 trained models, and a score of 0.527 for the final trained model. The topics produced by this model were then labeled by applying the same procedure as described in the Modeling section for labeling the LDA topics. The label and the top 10 terms for each topic are shown in Table 7, as well as the number of tokens assigned to each topic.

The number of tokens assigned to each topic gives an indication of the size and importance of each topic. Each of the topics were organised into one of seven categories. These categories are: (1) features, (2) connectivity, (3) acquisition, (4) workouts, (5) physical design, (6) physiological monitoring, and (7) miscellaneous. Figure 2 shows the seven categories and the relative size of each category.

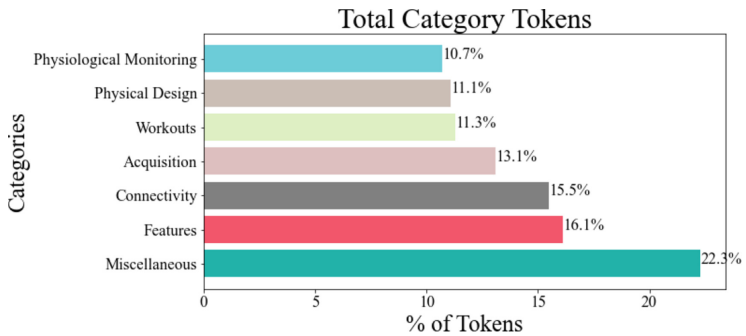

Fig. 2. Categories of topics and the number of tokens assigned to each category.

Table 7. Top terms and number of tokens assigned to each topic.

Label	Top 10 Terms	% of Tokens
Junk1	long, end, hope, case, post, wonder, couple, okay, similar, let	4.76%
Information	start, delete, check, fine, idea, late, link, numb, user, website	4.70%
Customisation Features	datum, way, add, option, miss, edit, amp, save, possible, available	4.66%
Junk2	pretty, bite, little, reason, actually, definitely, hard, easy, feel, expect	4.56%
Purchasing	wearable3, new, year, buy, month, old, wait, release, return, replace	4.56%
Junk3	day, week, change, hour, minute, today, notice, normal, lose, yesterday	4.48%
Accuracy	time, read, different, tell, guess, experience, correct, close, assume, example	4.46%
Junk4	wearable2, come, right, far, switch, leave, sound, live, reply, comment	4.29%
Settings	phone, set, turn, button, bluetooth, setting, stop, pair, mode, connection	4.24%
Junk5	point, mean, base, stuff, understand, basically, useful, information, real, kind	4.19%
Size/Weight	wear, wrist, big, small, compare, feel, vs, light, size, fit	4.15%
Syncing	app, connect, sync, track_app, os, download, competitor1_phone, create, upload, file	4.13%
Updates	update, notification, support, send, remove, software, version, firmware, contact, install	4.02%
Statistics	run, cycle, metric, test, load, drop, fast, long, min, short	3.94%
Upgrades	wearable1, plus, price, wearable8, model, upgrade, series, worth, deal, pro	3.79%
Logging	activity, step, record, walk, count, exercise, calory, goal, log, active	3.70%
VO_2-max	accurate, high, low, weight, vo_max, rest, accuracy, measure, average, level	3.68%
Water Features	feature, swim, fitness, sport, wearable10, prefer, runner, decide, water, touchscreen	3.67%
Heart Rate	hr, heart_rate, strap, sensor, hrm, monitor, chest_strap, wearable4, optical, ant	3.63%
Planning	workout, train, pace, plan, distance, mile, zone, lap, treadmill, race	3.62%
Watch Band	band, online_store, buy, sapphire, scratch, black, sell, mm, sale, discount	3.47%
Watchface	screen, battery, charge, face, battery_life, display, widget, field, power, wearable11	3.45%
General Health	track, sleep, competitor2, bike, ride, night, wearable7, stress, tracker, body_battery	3.39%
Geolocation	gps, map, hike, course, route, elevation, trail, weather, location, altimeter	3.31%
Audio	music, competitor1, pay, smart, play, spotify, competitor1_wearable, control, wifi, smartwatch	3.14%

The features category is, except for the miscellaneous category, the largest, and constitutes 16.1% of all the tokens in the corpus. It consists of the customisation, accuracy, water features, and geolocation topics. These topics are an indication of what kind of features consumers are interested in. The customisation topic suggests that consumers expect a large degree of control over how their wearable device functions and how they interact with it. Words like 'possible', 'available', 'miss', and 'option' could point to users' increased interest in finding out if such options do exist. The accuracy topic suggests that consumers sometimes find that readings do not match expectations or are not logical and thus not accurate. Closer inspection of representative documents also seem to suggest a keen interest in how certain measurements are obtained or how certain metrics are calculated. The water features topic is mainly concerned with how the wearable device behaves in water with regards to touchscreen sensitivity. The geolocation topic suggests that consumers often use their devices in the outdoors and use a wide variety of sensors, such as altimeters, barometers, and temperature sensors to monitor their environment.

The software category constitutes 15.5% of all the tokens in the corpus, and consists of the audio, device connection, syncing, and updates topics. These topics are all related to software and connectivity, and inspection of these topics' representative documents suggest that issues in this regard are numerous. The audio topic is mainly related to controlling audio feedback (mainly music) on other devices through the user's wearable. The top words also contain other brand names, indicating which other devices or music apps the wearables are most often used alongside. The settings topic is related to how users can access and change different features and modes on their devices. Similar to the customisation topic, the top terms suggest that having control over how the wearable functions is very important to consumers. The syncing topic is related to the data that is shared between different apps, devices, and platforms, and how users interact with this data. The top terms and representative documents signal the use of multiple apps and devices for daily activities, and that it is expected that these must be able to communicate and transfer data seamlessly. The updates topic concerns the continued software support and enhancements that Brand Y's devices receive.

The acquisition category consists of the information seeking, upgrades, and purchasing topics, and 13.1% of all tokens are assigned to this category. The information seeking topic is about where users find information and the different resources that they consult when seeking help or making a decision. The upgrades topic revolves around product offers and new versions of devices, and whether these compare well with what users already own. It also seems that these users actively look for discounts and deals when considering an upgrade. The purchasing topic is closely related to the upgrade topic, with users comparing different models and deciding whether to buy now or wait, but this topic also concerns issues such as shipping and availability of different models in different countries.

The workouts category constitutes 11.3% of all tokens in the corpus, and contains the statistics, logging, and planning topics. The planning topic's top terms indicate that being able to plan multiple aspects of a workout (such as distance, sequence, and pace) and comparing them with actual performance is important to users. The logging topic suggests that setting goals and logging activities that contribute towards those goals is also important to users. These goals also include special rewards or badges for completing challenges designed by Brand Y. The statistics topic indicates that users increasingly analyse their own fitness data and are interested in trends in their health and fitness.

The physical design category constitutes the watch band, watch face, and size/weight topics. The watch face topic seems to mainly focus on the screen and displays of watches, and their effect on the battery life of the watch. The watch band topic seems mainly concerned with the buying and selling of watch bands of different colours and materials. The size/weight topic reflects the importance of comfort and fit and how the size and weight of the watch affects this.

The physiological monitoring category consists of 10.7% of the tokens and, it includes the general health, heart rate, and VO_2-max topics. The general health topic is focused on the daily habits and lifestyle patterns that the wearables are able to track, such as sleeping, heart rate variability, and stress. The heart rate topic signals the importance of this single measurement to users, and the representative documents suggest a keen interest in its accuracy. The VO_2-max is also indicative of this specific metric's importance to users.

All junk topics were assigned to the miscellaneous category, since it is difficult to discern any single concept related to wearables from these topics' top terms. The top terms include 'wonder' and 'hope' in junk 1, 'actually' and 'expect' in junk 2, 'today' and 'notice' in junk 3, 'far' and 'come' in junk 4 and 'understand' and 'useful' in junk 5. These are all terms typically found in discussion forums where users share their experiences and opinions with others, and so it is not surprising that these words often occur together and appear as top terms for this specific dataset. The more emotive and descriptive nature of these words, however, mean that they do not add much value in terms of identifying specific aspects of products that are important to consumers. Additional cleaning and more extensive inspection of the corpus and vocabulary could prevent the formation of such junk topics.

7 Conclusion

This paper explored the influence and impact of topic modeling on market research practice and tested the performance of three topic modeling techniques as applied to a social media dataset extracted from Reddit, within the context of market research in the fitness wearable industry. An optimised LDA model with $k = 25$ was used as the baseline against which to measure the CTM and GSDMM algorithms. GSDMM provided worse results than the LDA, while CTM provided better results. The 25 topics produced by the CTM were then labeled by inspecting each topic's top 20 terms and its most representative documents.

Twenty of these topics were easily interpreted and received appropriate labels, while no appropriate labels could be found for the remaining five topics, and they were consequently labeled as 'junk'. The labeled topics were found to fall within larger categories, and each category was interpreted. The top terms for each topic gives an indication of *what* is important to customers, while the number of tokens belonging to each category provide information on the *degree* of importance of each category. There also exists the opportunity to use the topic categories to perform customer or product segmentation. The appearance of competitor brand names or their products in the top terms of topics may also provide an indication of the company's largest competitors from the consumers point of view. Perhaps the greatest advantage of using topic modeling for market research is the automatic and time-efficient extraction of topics from large corpora of social media text, which would otherwise require an exorbitant amount of time or manpower. The observational nature of the research may lead to the formulation of further research questions and provide direction for more focused and in-depth market research. The human analyst plays an important role in interpreting, labeling and categorising the topics, and in making the final decision about the number of topics that provides the best insight.

Future research may include the optimisation of hyperparameters for both CTM models and the GSDMM models, which includes finding the optimal number of topics and concentration parameters. In addition, added cleaning steps to remove semantically meaningless words could prevent the formation of a large number of junk topics.

Acknowledgements. This work is based on the research supported in part by the National Research Foundation of South Africa (Grant number: 129340), and the school for Data Science and Computational Thinking at Stellenbosch University.

References

1. Aggarwal, C.C., Zhai, C.: Mining Text Data. Springer, New York (2012). https://doi.org/10.1007/978-1-4614-3223-4
2. Asuncion, A., Welling, M., Smyth, P., Teh, Y.W.: On smoothing and inference for topic models. In: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009, pp. 27–34 (2009)
3. Batrinca, B., Treleaven, P.C.: Social media analytics: a survey of techniques, tools and platforms. AI Soc. **30**(1), 89–116 (2015). https://doi.org/10.1007/s00146-014-0549-4
4. Bendle, N.T., Wang, X.S.: Uncovering the message from the mess of big data. Bus. Horiz. **59**(1), 115–124 (2016)
5. Blei, D., Lafferty, J.: Correlated topic models. In: Advances in Neural Information Processing Systems, vol. 18, p. 147 (2006)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
7. Boyd-Graber, J., Mimno, D., Newman, D.: Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements. CRC Press, Boca Raton (2014)

8. Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: Neural Information Processing Systems, vol. 22, pp. 288–296. Citeseer (2009)
9. Chernova, V., Starostin, V.: Leading trends in marketing research: a case of big data. In: 5th International Conference on Social, Economic, and Academic Leadership (ICSEALV 2019), pp. 210–215. Atlantis Press (2020)
10. Cluley, R., Green, W., Owen, R.: The changing role of the marketing researcher in the age of digital technology: practitioner perspectives on the digitization of marketing research. Int. J. Mark. Res. **62**(1), 27–42 (2020)
11. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108 (2010)
12. Nunan, D., Di Domenico, M.: Market research and the ethics of big data. Int. J. Mark. Res. **55**(4), 505–520 (2013)
13. Patino, A., Pitta, D.A., Quinones, R.: Social media's emerging importance in market research. J. Consum. Mark. **29**(3), 233–237 (2012)
14. Qiang, J., Qian, Z., Li, Y., Yuan, Y., Wu, X.: Short text topic modeling techniques, applications, and performance: a survey. IEEE Trans. Knowl. Data Eng. **34**(3), 1427–1445 (2020)
15. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408 (2015)
16. Yin, J., Wang, J.: A dirichlet multinomial mixture model-based approach for short text clustering. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 233–242 (2014)