IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 9, 2022, accepted March 26, 2022, date of publication April 5, 2022, date of current version April 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3165043

Using Machine Learning to Improve Lead Times in the Identification of Emerging Customer Needs

DAVID KILROY[1], **GRAHAM HEALY**[2], **AND SIMON CATON**[1]
[1]School of Computer Science, University College Dublin, Dublin 4, D04 N2E5 Ireland
[2]School of Computing, Dublin City University, Dublin 9, D09 Y074 Ireland

Corresponding author: David Kilroy (david.kilroy1@ucdconnect.ie)

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

ABSTRACT In recent years, computational approaches for automatically extracting the voice of the customer from user generated content have been proposed. These studies have tackled the task of obtaining current customer needs, however, there is a lack of methods that predict future needs (i.e. needs that may become popular in the marketplace). Therefore, this study presents a multi-document keyphrase extraction algorithm which predicts future customer needs from users' social media posts on Reddit. Key to our approach is a novel document filtering method (discovering potentially relevant social media content) and a keyphrase ranking method, which promotes terms with rising frequency likely to be future product needs. In order to evaluate the approach, a case study of ''toothpaste'' needs is reviewed and a novel evaluation approach using ground truth automatically extracted from a collection of future specifications of new-to-market products is proposed. In our evaluation, we show that the approach is significantly better than simple baselines at identifying customer needs on social media before they trend in the marketplace. We also show that our approach can capture important customer needs identified by a large multinational company with lead times of up to 25 months ahead of them trending in the marketplace. The findings of this research could provide many benefits to businesses such as gaining early access into markets ahead of their competitors and giving early notice to manufacturers/engineers/developers before a need for a product is in demand.

INDEX TERMS Machine learning, product development, Reddit, text mining.

I. INTRODUCTION

The creation of new products has been defined as one of the ''top goals'' [1] as well as a ''critical success factor'' of a business [2] trying to grow or survive. The process of generating new product ideas or features for products has been defined as the ''product discovery'' [3], [4] or the ''new product development'' phase [5], [6]. Listening to the Voice of the Customer (VOC) when generating new ideas is critical [3], [7], [8], with customer satisfaction being ''vital'' [5] for the success of new product ideas. Some research has even pointed out that customers on their own can provide better product ideas than professional developers, with [9] finding that customers are likely to produce more novel products which earn more sales and [8] finding that customers provide better (albeit less feasible) product ideas.

The use of ''user interviews'' [10] is the most popular method for businesses to garner insights on their customers' needs and requirements. However, in recent years companies have also been turning to the use of statistical methods from Artificial Intelligence (AI), and in particular Machine Learning (ML), to gather requirements from User Generated Content (UGC) [11] such as product reviews [12]–[22] or social media [23]–[30]. These methods have warranted exploration because user interviews have drawbacks that computational methods don't suffer from, namely: 1) they incur large monetary and time costs; 2) they suffer from small sample sizes; and 3) they have the potential for biased results due to participants already knowing what responses the company wants [28]. ML methods that run over UGC are in such demand that they have become businesses within themselves, with companies such as Sprinklr, Meltwater and Sprout Social all engaging in the practice. The billion dollar valuation of Sprinklr, giving the company a ''unicorn'' status, shows the interest by other businesses in the area.[1]

The associate editor coordinating the review of this manuscript and approving it for publication was Wenbing Zhao.

[1] https://fortune.com/2015/03/31/sprinklr-unicorn-valuation/ - last accessed 01/04/2022

Researched approaches in the area of analysing UGC mine "customer needs", which has previously been defined in the marketing literature as "a description in the customer's own words of the benefit to be fulfilled by the product or service" [31] e.g. the need of *hydration* for the product Fanta. Computational approaches analysing UGC have also included the features or attributes of a product in this definition of "customer needs", as they express needs indirectly [32], [33] e.g. the need of *eco-friendly* for the *paper straw* that accompanied the Fanta product. Approaches which mine these needs are effective at analysing them from static collections of documents. However, as pointed out in a recent study [25], these methods fail to analyse needs over time which consequently hinders their effective use in business contexts that require constant timely feedback in order to make products or new product features that are required by customers. Aside from the limitation of previous studies not analysing customer needs over time, a lot of these studies also don't address the task of predicting the future importance of their customers' needs, as pointed out in [34]. Performing this task is important in real-world business scenarios as there are benefits associated with producing new products based on customer needs which are expected to exist in the future [35], such as gaining an advantage over competitors or reducing the need for future design changes [36].

A key hurdle in achieving the aforementioned aims has been a lack of intelligent keyphrase extraction algorithms for predicting future customer needs. Therefore, in this paper we propose two keyphrase extraction algorithms for obtaining customer needs from two diverse collections of text documents. The first algorithm uses a dataset of timestamped social media posts from Reddit while the second algorithm uses a dataset of timestamped (by release date) product descriptions, which postdates the social media data. By employing our novel approach we test whether our intelligent keyphrase detection algorithm run over the social media posts can predict the trending keyphrases appearing in the set of future product descriptions, i.e. to act as early indicators of future needs. Although this is a challenging task with a presumed high degree of error, we show that our proposed approach can do this effectively with precision and recall scores significantly better than our defined simple baselines. We also demonstrate that our approach is effective at detecting disruptive new product ideas that are identified by a large multinational company with both good recall and considerable lead times (years), ahead of the ideas first trending on the market i.e. in our product descriptions data. Our approach thus addresses the aforementioned limitations by analyzing customer needs longitudinally over multiple time windows and by predicting their future needs.

There are four unique contributions of this work:

- In contrast to previous approaches that mine needs from static collections of documents with no associated timestamps, our approach analyses these needs over time which is more more aligned with real-world business environments where customer requirements are needed on a consistent basis.
- The challenging task of predicting the future importance of customer needs is addressed, which would allow a company to obtain needs ahead of their competitors.
- A data reduction method to discover more relevant social media posts on Reddit for the task of identifying future customer needs is proposed.
- A novel evaluation strategy for the prediction of future customer needs is proposed (modelled closely on proven evaluation approaches from text mining).

The remainder of the paper is organized as follows. Section 2 provides a literature review of the related approaches for extracting customer needs from UGC. Section 3 illustrates the proposed method. Section 4 discusses the proposed evaluation, provides results and demonstrates our approach on finding future oral care needs. Finally, Section 5 concludes the study and discusses future research directions.

II. RELATED WORK

Studies of novel automated systems to extract customer needs from UGC have mainly relied on using techniques from text mining, natural language processing and ML [37]. These studies can be categorized into various distinct groups based on 1) the data used (Section II-A); 2) the application scenarios and methods performed (Section II-B); and 3) the evaluation strategy conducted (Section II-C). The rest of this section discusses these studies regarding the mentioned factors.

A. DATA USED

The types of data used for extracting customer needs from UGC mainly consist of social media [23]–[30] and product reviews [12]–[22]. Social media has the drawback of containing a large number of posts which are irrelevant to customer needs when compared to product reviews e.g. users discussing the product "mobile phone" for reasons other than needs they desire for it - "I texted them back on my mobile phone". However, given the context of our research aiming to discover future needs, social media may be more suitable. This is because product reviews are more likely to discuss customer's current needs compared to social media which has been used as a test bed for new and emerging needs and ideas e.g. social media users discussing DIY solutions to beauty products before they've become popularized in the market [38].

We select the social media platform Reddit for our analysis due to it being one of the few open opinion-based Application Programming Interface (API) platforms since the "Post-API Age" [39] of platforms like Twitter and Facebook restricting access following major data scandals e.g. Facebook-Cambridge Analytica [40]. Specifically, when obtaining data we use the Pushshift API, which has an API limit "five times greater" [41] than the official Reddit API further facilitating the data acquisition process. Baring data access/acquisition, previous research using Reddit for customer needs mining [23]–[25] has noted the advantages of using the platform. Previous research has stated the benefit of having data organized into defined ‘‘subreddits’’ (discussion forms) when capturing needs where platforms like Facebook, Twitter and Instagram are limited in this sense [23] (e.g. the subreddit r/PickAnAndroidForMe for Android phone products). [23], [25] also show that documents are generally longer in Reddit than in other platforms (e.g. Twitter’s 280 character limit), which could potentially help in mitigating the short text problem [42], [43] if a technique like clustering is performed over a collection of posts.

The predictive power of Google Trends has been used before in many business intelligence studies, most notably in sales prediction [47], [48]. More recently however, it has been used in order to estimate the importance of future customer needs, with [49] showing that it can be a predictive data source when applied in conjunction with product reviews. Similarly, our approach uses Google Trends alongside Reddit when ranking keyphrases representing customer needs of future importance.

B. APPLICATION SCENARIOS AND METHODS PERFORMED

Previous approaches in the literature that mine customer needs from UGC mainly follow the same general process. As discussed, the data is first collected, which may relate to an exact product [17], [23], [24] (e.g. iPhone 5) or a group of products [25], [28] (e.g. cell phones). Secondly, an amount of filtering or preprocessing of the data is carried out. Such filtering may consist of removing uninformative reviews or social media posts from the set of all documents [45], while preprocessing usually consists of the tokenization, normalization (e.g. stemming or lemmatization [14], [27], [28]) and removal of uninformative words from the documents themselves (e.g. stop words [25], [45]). Finally, an analysis is carried out using methods that fit the task to be solved. A subset of these tasks and methods will be discussed in this section.

Before these tasks and methods are discussed, techniques which filter the initial set of uninformative reviews or social media posts from the literature will be examined. A large body of work (not directly related to the extraction of needs from UGC) have focused on classifying documents based on ‘‘purchasing intent’’ (i.e. desire to buy a product) [50]–[52]. These methods have been used on manually labelled data from a variety of different platforms e.g. Quora [50], Yahoo Answers [50] and Twitter [51], [52]. Work in the field of customer needs mining has borrowed these techniques in order to reduce the number of documents for analysis before extracting customer needs [45]. Similarly, our approach works on a reduced set of documents by performing a novel data reduction technique specific to the platform Reddit, by only considering posts coming from subreddits which are similar to a subreddit which is known to discuss needs relating to a given input product (e.g. r/OfficeChairs for chairs). In our evaluation, we show how this reduction technique can lead to significantly increased results compared to when no reduction is performed. Our reduction technique could be considered complementary to the existing ‘‘purchasing intent’’ literature as it does not remove posts in the same manner, thus potentially removing uninformative posts which wouldn’t have been found otherwise.

A number of authors have investigated the extraction of customer needs by focusing on ‘‘product opportunities’’ [23]–[25], [53]. In [24], opportunities are identified through the theory of ‘‘Chance Discovery’’, initially proposed in [54], in which analysts are presented with a keygraph in order to identify customer needs. In the keygraph, the keys are documents and the documents’ similarity to each other is primarily computed from the output of the document-topic matrix returned by Latent Dirichlet Allocation (LDA). In [23], [25], the concept of an ‘‘opportunity algorithm’’ is implemented with respect to finding opportunities for new or existing customer needs. As initially described in [55], this algorithm works on the basis that if a customer need has high importance but low satisfaction then a business opportunity is present. In [23], [25], these importance and satisfaction values are computed based on the degree to which a customer need is discussed (importance) along with the sentiment of each of these needs (satisfaction). In [23], this is applied to the outputs of the term-topic matrix returned by LDA, where the goal of the analysis is to find product opportunities for the mobile phone ‘‘Galaxy Note 5’’. Although there has been considerable work on tracking events on social media over time [25], [56]–[58] noted the lack of methods in tracking product opportunities on social media over time and proposed an aging-based Event Detection and Tracking (EDT) clustering model, where the goal of the analysis is to track events of customer needs and observe whether they could potentially become product opportunities by applying the ‘‘opportunity algorithm’’. The proposed EDT model analyzes the lifecycle of a customer need by observing its birth, growth, decay and eventual death. Although our approach does not analyze customer needs in this way (i.e. proposing a detection and tracking clustering model), it presents similar contributions to [25] by analyzing customer needs over time by running a keyphrase extraction algorithm that produces lists of ranked customer needs in fixed time windows.

Similarly to the implementation of the ‘‘opportunity algorithm’’, additional lines of research have also focused on applying text mining and ML techniques to popular business models or methodologies, with Kansei engineering and the Kano model being two key examples. Kansei engineering can be defined as ‘‘translating technology of a consumer’s feeling and image for a product into design elements’’ [59]. Traditional approaches for Kansei engineering work on questionnaires to measure a user’s feelings towards a customer need where groups of words called ‘‘Kansei attributes’’ are used to measure their emotions. Recent computational approaches try to implement Kansei engineering on UGC so that time-consuming questionnaires don’t need to be carried

TABLE 1. Summary of studies for customer needs mining.

Study	Purpose	Data	Methodology	Predict Future Needs	Analyze Needs Over Time	Perform Data Reduction	Evaluation on Real-World Importance
[24]	Identify product opportunities	Social Media (Reddit)	Keygraph & LDA	No	No	No	No
[23]	Identify product opportunities	Social Media (Reddit)	Opportunity Algorithm & LDA	No	No	No	No
[25]	Time-evolving product opportunities	Social Media (Reddit)	Aging Based Event Detection & Tracking	No	Yes	No	No
[44]	Feature-Affective Opinion Extraction	Product Reviews (Amazon)	Text Mining Techniques	No	No	No	Yes
[45]	Customer need (dis)satisfaction	Product Reviews (Amazon)	LDA & Sentiment Analysis	No	No	Yes	No
[46]	Predicting future customer needs	Product Reviews	Holt-Winters Exponential Smoothing	Yes	Yes	No	No
[34]	Predicting future customer needs	Product Reviews (Amazon)	Fuzzy Time Series	Yes	Yes	No	No
our approach	Predicting important customer needs over time	Social Media (Reddit) & Google Trends	Keyphrase Extraction & MK Test	Yes	Yes	Yes	Yes

out [13], [14], [44], [60], [61]. [44] extracts customer needs from Amazon reviews using the linguistic information of words and then expands on the list of Kansei attributes (first identified by numerous research works in the product development literature) using WordNet. In their approach, they show that emotions towards a customer need in "feature-affective" pairs can be extracted with "high precision and recall" [44]. The Kano model is a product development theory to weighting customer needs to the extent in which they satisfy/dissatisfy customers [62]. Similarly, computational approaches are used to implement this model [12], [16], [45], [63], [64]. For example, the approach in [45] implements it by looking at the varying degrees of sentiment applied to product-topics returned by LDA in order to get the levels of satisfaction and dissatisfaction of customer needs. These business model approaches (i.e. Kensei and Kano) are effective at extracting needs from customers. However, they don't address the contributions defined in this research, as they mine on static collections of documents and don't address the task of the future prediction of needs.

[26], [34], [46] tackle the problem of predicting the future importance of customer needs. In [46], needs from customer reviews of cell phones are extracted and then predicted using Holt-Winters exponential smoothing. In their evaluation, they show how features such as "battery life" are predicted over time. In [34], a fuzzy time series model is applied to predict the future importance (based on frequency and sentiment) of customer needs from Amazon reviews of electric iron products. In their experiments, they compare their approach to other time series models and show that their fuzzy time series models are able to predict the importance of customer needs better than other classical time series models (e.g. Simple Moving Average). Similarly to the mentioned studies, our approach also predicts the future importance of customer needs. It does this by applying techniques from trend analysis in order to determine the degree of increase in a keyphrase's usage over time by considering data from two sources (Reddit and Google Trends). However, unlike these studies our approach addresses the prediction problem differently by instead producing ranked lists of future customer needs in the form of keyphrases for a given time window instead of predicting the importance of individual needs as a regression problem. Hence, the strategy used to evaluate our approach is adapted to address this problem.

C. RESEARCHED EVALUATION STRATEGIES CONDUCTED

Strategies that have been proposed for the evaluation of customer needs from UGC are usually dependent on the problem that is being addressed. For example, [44] compare their algorithm's performance at extracting feature-affective pairs to manually labelled instances from 10 products out of the "Toys & Games" category on Amazon. [26], [27] assesses the effectiveness of their algorithms at categorizing customer needs into strong, weak and controversial features for various selected smartphones [26], [27] and automobiles [27].

Approaches that forecast future customer needs often apply regression techniques in order to predict an "importance score" [34], [46] e.g. in [34] this "importance score" is based on the frequency and sentiment of a need. These techniques then apply regression metrics (e.g. Prediction Error [34] or Mean Absolute Percentage Error [46]) in order to evaluate the prediction of their "importance scores". A problem with these approaches is that they don't correlate their defined "importance score" to a real-world value of importance. An example of an unrelated task which does

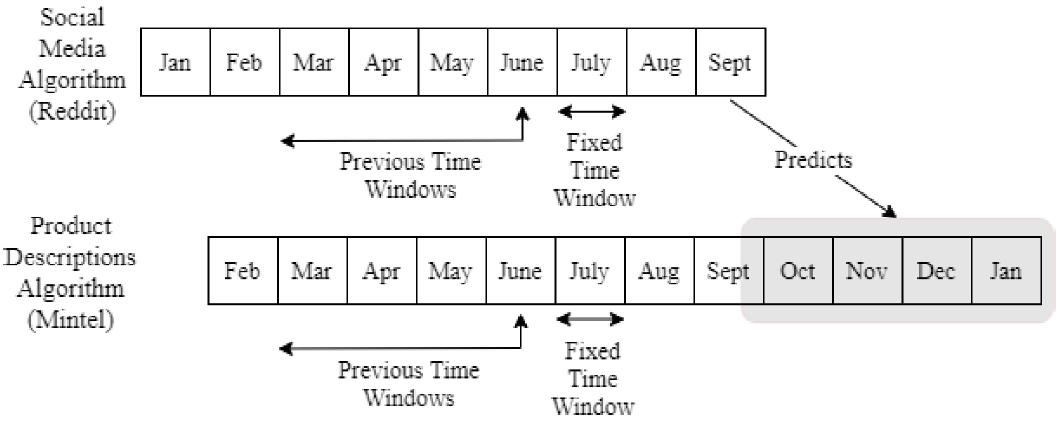

FIGURE 1. Overview of task (social media predicts trending keyphrases in product descriptions).

this is in [65], who correlate the "chatter" for a movie on Twitter to actual box office revenue. In our approach, we attempt to define a real-world value of importance for customer needs based on needs occurring in future product descriptions. As such, we require different metrics to evaluate our approach.

Table 1 summarizes the key studies mentioned in this section and shows the advantages of our approach and where it fills the gap in the research when compared to other studies. The table shows that contributions are made in 1) the prediction of future needs; 2) the analysis of needs over time; 3) removal of uninformative documents (i.e. data reduction); and 4) the evaluation of customer needs on real-world importance scores. Through the aforementioned contributions, our work would allow product development teams to have a framework for finding new and emerging customer needs that could then be addressed in the products they make.

III. PROPOSED APPROACH

Figure 1 outlines the keyphrase prediction problem we address in this study. In brief, the aim of the proposed approach is to extract keyphrases from social media which predict keyphrases representing customer needs in product descriptions as far into the future as possible. In order to perform this task, we apply two algorithms: a) social media algorithm (which runs over textual social media posts) and b) product descriptions algorithm (which runs over textual product descriptions). Both of these algorithms make use of the timestamp associated with their retrospective documents (i.e. social media post or product description) in order to produce a ranked list of keyphrases at each fixed time window e.g. produce keyphrases for each month (as in Figure 1).[2] In order for both of the algorithms to produce these ranked keyphrases at each fixed time window, they consider data from previous time windows. This is seen in Figure 1, where both of the algorithms use data from 3 previous time windows (i.e. April, May and June) in order to produce their final ranked keyphrase list for an individual fixed time window (i.e. July). Data from these previous time windows is required due to the nature of the methods being used by each of the algorithms (requiring past data for their computation).

In our experiments, the product descriptions algorithm is used to generate the ground truth for evaluating the social media algorithm's outputs i.e. top keyphrases at each fixed time window, which represent new and emerging needs in products entering the market at that time. The social media algorithm (run over data from Reddit) is then optimized to identify the important keyphrases which will appear in the future top keyphrase lists produced by the product descriptions algorithm. The overall task is then to observe whether there is a future time lagged relationship between the keyphrases produced by the product descriptions algorithm and the social media algorithm. This is seen in Figure 1 where the keyphrases generated in the fixed time window by the social media algorithm (i.e. September) aim to predict the keyphrases produced 1 - 4 months in the respective future by the product descriptions algorithm (i.e. October, November, December and January). This also enables the measurement of how far in advance social media data can predict customer needs identified by important keyphrases (e.g. product features) that will occur in future products. It is important to note that the methodology used to produce keyphrases for the social media algorithm does not require the keyphrases from the product description algorithm, rather it's only used to evaluate it.

Table 2 shows the key parameters used in the approach. Some of these have already been discussed (e.g. fixed time window length) while the rest will be detailed throughout the remainder of this section. We discuss and present each parameter in this section without reference to "good" or "suitable" values, instead, refining the values of many parameters forms a key part of our evaluation (Section IV-B).

Figure 2 outlines each of the algorithm's approaches when producing ranked keyphrases for each fixed time window. The key components seen in the figure make up the

[2] The exact time frame seen in Figure 1 is not the one used in the experimental setup but is rather used to illustrate how the task is performed.

TABLE 2. Description of parameters used in the methodology for the product description and social media algorithms.

Parameter Name	Parameter Type	Description
Fixed Time Window Length	Experimental Setup	The time span in which to produce customer needs
Num. Past Time Windows	Experimental Setup	The number of past time windows of data to use in order to produce needs at each fixed time window
Future Prediction Time	Experimental Setup	The defined window of time the social media algorithm tries to predict needs in future product descriptions
Pdesc Min Document Frequency	Product Descriptions Algorithm	The min document frequency of a keyphrase in a set of descriptions in order for it to be considered a need
Social Media Target Keyphrase	Social Media Algorithm	A keyphrase which adheres to a target product in order to collect posts on social media for the analysis
Google Trends Category	Social Media Algorithm	The google trend category which is related to the target product being analyzed
Gold Standard Subreddit	Social Media Algorithm	The subreddit which is related to the product under analysis, which is used by the data reduction approach
% Most Similar to Gold Standard Subreddit	Social Media Algorithm	A parameter value used to control the number of posts used in the analysis by including/excluding posts based on their similarity to the *Gold Standard Subreddit*
Allowed POS Tags	Social Media Algorithm	The allowed part-of-speech tags for a keyphrase to be considered a customer need
Social Media Min Document Frequency	Social Media Algorithm	The min document frequency of a keyphrase in a set of social media posts in order for it to be considered a need
Min Chi Square P-value	Social Media Algorithm	The min chi square value a keyphrase must have when its frequency on Reddit is compared to a reference corpus

subsections of this section: 1) the data used as input into each of the algorithms (Section III-A); 2) the algorithm run over the product descriptions data (Section III-B); and 3) the algorithm run over the social media data (Section III-C). This section only focuses on providing an overview of the methods used in our approach. A specific case study example of ‘‘toothpaste’’ customer needs extraction will be discussed in Section IV-B.

A. DATASETS

1) SOCIAL MEDIA DATA

The social media dataset used in this study is from Reddit. Past approaches using Reddit data for the purposes of mining customer needs have looked at specific subreddits in order to extract needs for a particular product type [23]–[25] e.g. r/chairs for chair products. As in [28], instead of looking for posts on a specific subreddit, our approach instead searches for posts with a target keyword that the product is associated with e.g. for chairs, only posts with the keyword ‘‘chair’’ are searched for (*Social Media Target Keyphrase* - Table 2). Furthermore, our approach doesn’t consider data from all subreddits. It does this by applying a filtering technique in order to only consider data from subreddits which are likely to discuss customer needs relating to a product (discussed later in this section).

2) GOOGLE TRENDS

Recently Google Trends has been used in studies in order to predict future customer needs [49]. We also use this data source in conjunction with Reddit when producing a final ranking of customer needs for a given corresponding time window. It returns a time series for a particular search term indicating the search volume of different queries over time. Given that this time series is representative of a larger group of people (compared to Reddit), means it gives a good measure of the importance of a particular customer need at a given time period. Google Trends also allows for the ‘‘category’’ of search to be selected when searching for a keyphrase, thus allowing keyphrases to be searched for with respect to a given input category e.g. ‘‘grapes’’ in the ‘‘Non-Alcoholic Beverages’’ category (*Google Trends Category* - Table 2).[3] This parameter can be informative when observing how important a need is with respect to a product category (e.g. how important is ‘‘grapes’’ in the ‘‘Non-Alcoholic Beverages’’ product category) rather than just the entire body of searches (e.g. how important is ‘‘grapes’’ in all of Google’s searches).

3) PRODUCT DESCRIPTIONS DATA

The product descriptions data captures all product types that are searched for on social media e.g. if the task is to find needs for the product ‘‘hand lotion’’ on social media, then a dataset of product descriptions for ‘‘hand lotion’’ is required for the evaluation. In our study, we analyze product descriptions from Mintel Global New Products Database (GNPD), which is a large product information database used by businesses and academics that are involved in marketing and innovation [66]. GNPD analyzes various brands from more than 50 countries over a number of different product categories (e.g. food, drink, pet products, beauty, personal care etc.) [66]. The information that is available for each

[3] https://github.com/pat310/google-trends-api/wiki/Google-Trends-Categories – last accessed 01/04/2022

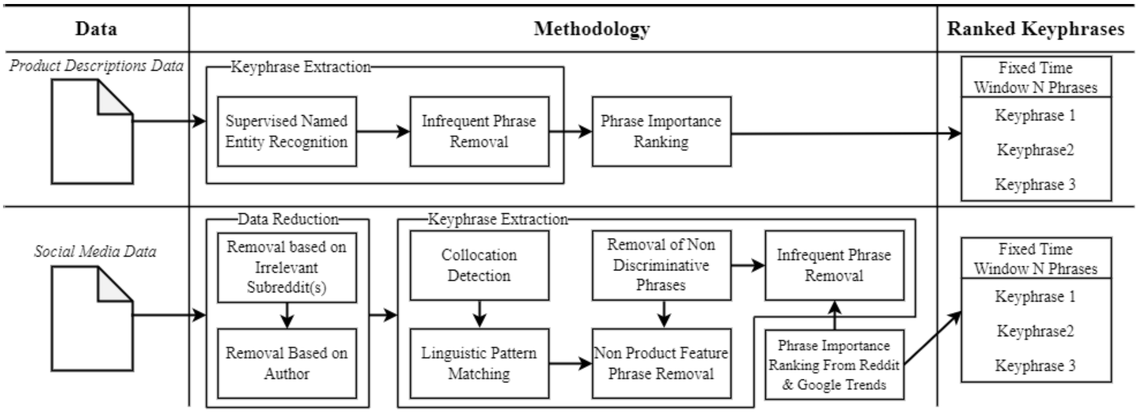
Data
Methodology
Ranked Keyphrases
Product Descriptions Data
Keyphrase Extraction
Supervised Named Entity Recognition
Infrequent Phrase Removal
Phrase Importance Ranking
Fixed Time Window N Phrases
Keyphrase 1
Keyphrase2
Keyphrase 3
Social Media Data
Data Reduction
Removal based on Irrelevant Subreddit(s)
Removal Based on Author
Keyphrase Extraction
Collocation Detection
Linguistic Pattern Matching
Removal of Non Discriminative Phrases
Non Product Feature Phrase Removal
Infrequent Phrase Removal
Phrase Importance Ranking From Reddit & Google Trends
Fixed Time Window N Phrases
Keyphrase 1
Keyphrase2
Keyphrase 3

FIGURE 2. Overview of the methodology for the production of ranked lists of keywords from product descriptions (ground truth) and social media (customer need predictions).

product is significantly detailed, with various descriptions discussing claims and product features which are found on the packaging of products which would in turn reflect the needs of customers [66]. These descriptions are manually written by Mintel after viewing the actual product, first collected by one of their "mystery shoppers".[4] Of specific relevance is the number of products contained in the database ($\approx$ 33,000 new products added each month [66]) as well as the fact that each product is timestamped with the date it is first made available for retail, thus allowing customer needs addressed in new products to be tracked accurately over time.

B. PRODUCT DESCRIPTIONS ALGORITHM

In this subsection, the product descriptions algorithm, which detects top trending kephrase needs each fixed time window for the evaluation of the social media algorithm, is introduced. It does this by using a Named Entity Recognition (NER) model in order to detect customer needs from product descriptions data which are then ranked based on how often they occur in the current window of interest compared to previous time windows. This is done with the goal of finding needs which are popular by showing bursts in keyphrase usage. As seen in Figure 2, the algorithm performs three main steps: 1) Supervised Named Entity Recognition; b) Infrequent Phrase Removal; and c) Phrase Importance Ranking. For the first step (Supervised Named Entity Recognition), three sub-steps are performed in order to extract the initial set of entities from the product descriptions: a) Annotation; b) Data Augmentation; and c) Run Ensemble of Supervised Models. All the mentioned steps and sub-steps will be discussed throughout the remainder of this subsection.

1) ANNOTATION (SUPERVISED NAMED ENTITY RECOGNITION)

For a given target product description, annotators label "customer needs", which is different to previous work on attribute extraction which focuses on entities such as the brand [67], [68] or the size/color of a product [68]. As in [32], [33], our definition of "customer needs" includes the benefiting specifications of the product as well as the features or attributes of a product which have benefits associated with them. In order to follow this definition, we split customer needs into two main categories: a) direct needs - directly stated benefits a user gets/overcomes from using a product (e.g., *"the fresh product is recyclable and helps in curing the users hay fever"* → {"fresh", "recyclable", "curing", "hay fever"}); b) indirect needs - features of a product which contain benefits (e.g., *"the product contains mahogany"* → {'mahogany'}). We use this definition as both sources of information are of interest to people who could be using this research (e.g. innovation teams who are interested in upcoming features that could be used in products).

Table 3 lists the main guidelines annotators follow when tagging these needs. In Rule 1, annotators only label entities as separate unigrams unless the pair/group of words is imperative to the meaning of the entity, e.g. *"the product contains tea tree oil"* → {"tea tree oil"}. This rule is applied because entities need to be grouped together and counted over time after running the NER algorithm. For negation cases (as in Rule 2), annotators only label the entity. This is followed for the same reasons as Rule 1, with many different forms of "does not contain" occurring before the main entity e.g. "free from", "excludes" etc. It could be argued to not tag these entities at all (as they are not in the product), however as it is explicitly mentioned in the description we felt that it is important to tag these entities. In reality, most of these types of entities included needs which would definitely not be a benefit to be included in a product, such as for environmental reasons (e.g. plastics) or health reasons (e.g. added sugar). Finally, rules 3-6 are applied as they are not needs relating to a product itself. Some of these rules could be argued as needs (e.g. color of a product), however, they are noted for labelling consistency purposes.

2) DATA AUGMENTATION (SUPERVISED NAMED ENTITY RECOGNITION)

The use of data augmentation has been proven to be effective for many language processing tasks e.g. text

[4] https://www.youtube.com/watch?v=VPNq2wO_g9o – last accessed 01/04/2022

classification [69], [70]. More recently, [71] showed that performing simple augmentation methods can significantly increase the accuracy for NER tasks, especially when the number of annotated samples is small. In light of this, our work makes use of all the 4 methods proposed in [71] in order to increase the size of the training data as well as the accuracy of the model. Three out of four of these methods are centered around substituting entities with similar ones (e.g. synonyms/entities with the same tag) while the final method deals with shuffling segments of text in order to augment the data. It's also important to note that each of the augmentation methods have two of their own hyper-parameters: 1) *number_of_additional_generated_instances* per annotated instance; and 2) the random probability of replacing a token for an annotated instance p.

3) RUN ENSEMBLE OF SUPERVISED MODELS (SUPERVISED NAMED ENTITY RECOGNITION)

Ensemble learning is a ML technique that uses multiple base ML models instead of only one model in order to increase prediction performance and stability [72]. Recently, ensemble learning has shown to be effective in NER tasks [73]–[75], especially when using neural models which can produce varying predictions from one set of model training to the next due to the element of randomness present in deep learning techniques [73]. In our approach, we also use ensemble learning due to the fact that our approach uses a neural model as well as a discussed augmentation technique containing an additional layer of randomness to it (i.e. the random probability of replacing a token for an annotated instance p). The base NER models we use in our approach are from the python natural language processing library spaCy. SpaCy uses the neural language processing framework ''Embed-Encode-Attend-Predict''.[5] In brief, each model works by embedding words into a vector representation (embed), then encoding word vectors into a sentence matrix which allows context to be accounted for (encode), after which the sentence matrix is converted into a single summary vector which allows for a more condensed representation (attend) before finally making a single prediction from the summary vector (predict). In our ensemble, we train a number of base models which each contain their own version of augmented data based on the original labelled data. The entities classified for each description are then picked based on a majority vote from all of the trained models (as in [73]). In our approach we choose the number of base models to be 19. The value picked could have been smaller (e.g. 3 base models) however we choose this high value in order to ensure the reliability of the entity extraction. After classification, each entity is lemmatized and converted to lowercase. These entities are then counted and grouped according to how often they occurred in the previous fixed time windows of interest, in order to form a *keyphrase:date* → *occurrence* accessor (e.g. feather: 2015-01-01→20). This transformation is performed for the next stages of analysis so that top trending entities could be found at each fixed time window of interest.[6]

4) INFREQUENT PHRASE REMOVAL

Phrases which didn't exceed a minimum document frequency in previous time windows of interest are removed as potential customer needs (*Min Document Frequency* - Table 2). This is done so that keyphrases which only occurred a low number of times wouldn't wrongly be declared as top trending needs.

5) PHRASE IMPORTANCE RANKING

The goal of the product descriptions algorithm is to extract customer needs which are trending at a specific point in time. Past approaches to this problem (in the area of topic detection) have tried to find what is currently popular by detecting ''bursts'' in activity [76], [77] using Kleinberg's burst detection algorithm [78]. Similarly, the statistical measure of a z-score has been used to detect ''bursts'' in keyphrase activity [79], [80]. In our approach, we use this measure of a z-score to rank the remaining candidate keyphrases. This is computed by taking into account the occurrence of a keyphrase at previous time windows as well as the current time window, computing the z-score and then extracting the z-score for the current time window of interest. These phrases are then ranked by their z-score value in the time window. The top number of ranked keyphrases are then used to represent the top customer needs in the time window of interest. It is important to note that companies develop new products in response to a perceived customer need in their target market(s). We cannot observe this process, and thus use the emergence of new trends in product descriptions as a proxy for customer needs.

C. SOCIAL MEDIA ALGORITHM

In this subsection we introduce the social media algorithm, which detects keyphrases likely to be future customer needs, i.e. keyphrases identified by the product descriptions algorithm. It does this by: 1) reducing the number of posts under analysis to a distilled set of posts which are more likely to discuss customer needs relating to a product (Data Reduction); 2) finding candidate keyphrases in the posts which are most likely to be about customer needs (Keyphrase Extraction); and 3) ranking keyphrases based off of how popular they are expected to be in the future (Phrase Importance Ranking). As seen in Figure 2, the algorithm performs eight main steps: 1) Removal of Posts Based on Irrelevant Subreddit(s) (Data Reduction); 2) Removal of Posts based on Author (Data Reduction); 3) Collocation Detection (Keyphrase Extraction); 4) Linguistic Pattern Matching (Keyphrase Extraction); 5) Non Customer

[5] https://explosion.ai/blog/deep-learning-formula-nlp – last accessed 01/04/2022

[6] As an aside, although we construct this algorithm for the generation of a set of ground truth observations for the evaluation of our social media algorithm, we note that this approach would also serve as a means of competitor intelligence in the evolution of competitors portfolio of products aligned to key user needs.

TABLE 3. Data annotation guidelines for labelling "customer needs" from product descriptions.

Rule No.	Rule	Example
1	Mainly tag entities as unigrams	The trolley contains a rechargeable battery → {"rechargeable", "battery"}
2	In a negation case, only tag the entity	The shampoo does not contain triclosan → {"triclosan"}
3	Do not tag any of the brand name or title	Tesco Luxury Strawberry FaceMask ... → { }
4	Do not tag additional products that are promoted alongside the product of interest	This shampoo comes free with the product → { }
5	Do not tag entities recommended by/targeted at a group of people	Doctors recommended this product for the elderly → { }
6	Do not tag the color of a product	This brown product ... → { }

Need Word Removal (Keyphrase Extraction); 6) Removal of Non-Discriminate Phrases (Keyphrase Extraction); 7) Infrequent Phrase Removal (Keyphrase Extraction) and 8) Phrase Importance Ranking From Reddit & Google Trends. All the mentioned steps will be discussed throughout the remainder of this subsection. Prior to any of these steps being performed, the standard techniques of tokenization, lemmatization and lowering are all performed.

1) REMOVAL OF POSTS BASED ON IRRELEVANT SUBREDDIT(s) (DATA REDUCTION)

With the collection of posts being distilled down to just posts which contain the keyphrase representing some product type (discussed in Section III-A), posts on specific subreddits have a low probability of discussing customer needs relating to a product type. An example of this is on r/Gaming where users often post about the keyphrase ''toothpaste'' when using the ''toothpaste method'' to clean their CD's, however rarely discuss toothpaste in terms of needs or features relating to a product.[7] In order to remove these subreddits, subreddits which are likely to post about a customer need are found by calculating their similarity to a *Gold Standard Subreddit* (Table 2). Similarly to how [81] found similar documents for the purposes of automated essay scoring, our approach finds similar subreddits to the defined *Gold Standard Subreddit*. This works by collapsing all posts from a specific subreddit into one document to arrive at a subreddit-term matrix, which is further transformed by turning it into a tf-idf representation. The closeness of each subreddit (document) is calculated by finding its cosine similarity to the *Gold Standard Subreddit*, which is known to discuss customer needs. The subreddits are then ranked in accordance to how similar they are to the *Gold Standard Subreddit* and the top percentile of subreddits are retained along with the posts they contain while the others are removed (*% Most Similar to Gold Standard Subreddit* - Table 2). In our evaluation we show how this data reduction approach is an important step in our methodology by showing that it can obtain statistically significant results at finding future customer needs compared to when no reduction is performed.

2) REMOVAL OF POSTS BASED ON AUTHOR (DATA REDUCTION)

Reddit moderators are removed as these users are primarily there to point out lapses in other users ''reddiquette'' [82] (i.e. an etiquette to follow whilst on Reddit), and hence do not discuss needs relating to products. Bots are also removed as they do not represent content from a real individual and also have a high probability of posting spam content [83], [84].

3) COLLOCATION DETECTION

After the number of posts is reduced, some words from the posts are grouped together if they collocate (have a strong tendency to occur side-by-side) e.g. ''tea tree oil'', ''baking soda'' etc. Work in the area of keyphrase extraction has looked at grouping words which collocate by having a high Normalized Pointwise Mutual Information (NPMI) score [85]. Similarly, our approach groups words into phrases if the words have a high NPMI score and also co-occur some minimum numbers of times.

4) LINGUISTIC PATTERN MATCHING

Various surveys in keyphrase extraction have noted the use of the linguistic properties of words when removing/including candidate phrases [86]–[88]. These surveys have pointed to the literature applying conditions on phrases such that they must contain a particular Part of Speech (POS) tag such as a noun when being considered. Similarly, our approach applies these restrictions by only allowing a candidate keyphrase to be selected if it has the POS tag of noun, adjective, verb or adverb (*Allowed POS Tags* - Table 2). These POS tags are chosen as they are the tags which we found to mainly discuss customer needs (discussed in Section IV-B).

5) NON CUSTOMER NEED WORD REMOVAL

Stop words, URLs and curse words (as in [89]) are removed due to having a low likelihood of being related to a customer need.

6) REMOVAL OF NON-DISCRIMINATE PHRASES

As with similar approaches which extract features from review data [90], [91], non-domain-dependent phrases are removed. These phrases are removed as they do not relate

[7] https://www.wikihow.com/Fix-a-Scratched-CD - last accessed 01/04/2022

to the needs of the product type being searched for. In order to find these phrases, we use the chi-square test [92] for the purposes of discovering if there is a statistically significant difference between the observed frequency of a phrase on Reddit to its expected frequency according to a large-scale reference corpus. Similarly to work in keyword extraction [93]–[95], the test is computed for each phrase using a 2-by-2 contingency table, in which a test statistic is returned along with its corresponding p-value. If the p-value associated with each phrase is above some set threshold (*Min Chi Square P-value* - Table 2), then the phrase is removed.

7) INFREQUENT PHRASE REMOVAL

As for the same reasons stated in Section III-B (Infrequent Phrase Removal), phrases which didn't exceed a minimum document frequency in the previous time window of interest are removed (*Social Media Min Document Frequency* - Table 2).

8) PHRASE IMPORTANCE RANKING FROM REDDIT & GOOGLE TRENDS

As the goal of the algorithm run over the social media data is to extract future customer needs, our approach applies a Mann Kendall (MK) trend test in order to estimate whether there is an increase in a keyphrase's usage over time. The MK trend test itself has been used in similar applications in analysing trends across several domains e.g. tracking participation trends [96] and analysing trends in scholarly articles [97] and social media [98]. For an individual candidate keyphrase the approach works by obtaining its time series from Reddit based off of data from its previous time windows. The normalized time series corresponding to the candidate keyphrase is then obtained from Google Trends.[8] The MK trend test is then run on both of these time series (Reddit and Google Trends). The slope values returned from each of these MK trend tests are then added together to get a final ranking value for a keyphrase in a time window. Prior to the series being inputted into the MK trend test, they are normalized using unit vector normalization. This is so that the slope from each platform is of equal weighting in the final ranking value and so that the slope value reflects the relative increase in a keyphrase rather than just a raw frequency increase. This ranking value therefore represents the keyphrase's increased usage and thus gives a measure of the future importance of the keyphrase. All the keyphrases are then sorted by this final ranking value and the top number of keyphrases are chosen as customer needs which are of future importance. In our evaluation, we detail the positive impact of the introduction of the Google Trends data and the way in which this ranking approach is performed (compared to if only the Reddit data is used to rank the keyphrases).

D. SUMMARY

In this section, the overview task of using social media data to generate keyphrases representing customer needs well in advance of their materialisation in product descriptions was given. In order to do this, we detailed two algorithms run over product descriptions and social media.

The goal of the product descriptions algorithm is to find representative keyphrase needs each fixed time window. It does this by applying an ensemble of NER models in order to extract customer needs from a described annotated dataset (Supervised Named Entity Recognition). After, it uses a simple statistical measure of a z-score in order to rank needs which are currently popular (Phrase Importance Ranking).

The goal of the social media algorithm is then to detect keyphrases which are likely to be future needs in product descriptions. It does this by first reducing the number of posts to a subset of posts which are likely to discuss customer needs relating to a product (Data Reduction) and presents a technique which removes posts based off of which subreddit they occurred in (Removal of Posts Based on Irrelevant Subreddit). It then applies various techniques from keyphrase extraction to remove/include candidate keyphrases as customer needs (Keyphrase Extraction). Finally, it ranks the candidate keyphrase needs using the MK trend test in order to find keyphrases which show an increase in usage over time by considering data from Reddit as well as Google Trends (Phrase Importance Ranking From Reddit & Google Trends).

IV. CASE STUDY & EVALUATION

This section aims to detail our solution with respect to an example case study while also seeking to answer the following two research questions: 1) can future customer needs (as identified in future product descriptions) be predicted using UGC from social media (Section IV-D and IV-E); and 2) is our data reduction approach (described in Section III-C, i.e. Removal of Posts Based on Irrelevant Subreddit(s)) useful for the task of predicting future customer needs (Section IV-F).

In order to show the effectiveness of our approach, we evaluate it on finding future customer needs for "toothpaste" products.[9] Although the toothpaste use-case may seem an unusual choice for the discovery of customer needs, we use it for two main reasons: 1) there are continually many new ingredients (e.g. charcoal [99]), and benefits (e.g. plant-based, disease prevention [100]), representing customer needs that are constantly being addressed in toothpaste products, thus making it an interesting test case to investigate whether these needs can be detected on social media before they trend in product descriptions; and 2) the broader area of oral-care is a multi-billion dollar global industry that is still growing, therefore companies creating toothpaste products

[8] https://support.google.com/trends/answer/4365533?hl=en - last accessed 01/04/2022

[9] The implementation and dataset for our solution with respect to finding future toothpaste customer needs can be found at - https://github.com/davidkilroy/Using-Machine-Learning-to-Improve-Lead-Times-in-the-Identification-of-Emerging-Customer-Needs

could benefit greatly from using this research.[10] It is also worthwhile noting that the oral-care sector has already been subject to research for identifying needs from UGC using computational and statistical techniques [101].

This section first describes the datasets we use in our experiments (Section IV-A). We then detail the steps of the methodology we apply for finding future customer needs for our defined case study of toothpaste needs (Section IV-B). The capability of our product descriptions algorithm is then accessed in order to observe how well it performs at extracting entities from product descriptions, thus examining how pure our ground truth data is (Section IV-C). Our experimental set-up is then explained and our social media algorithms performance at finding future customer needs in product descriptions is evaluated (Section IV-D). After, a case study evaluation is presented comparing our social media algorithm's performance on needs obtained from a large multinational company (Section IV-E). The impact of key steps in our methodology are detailed (Section IV-F). Finally, a summary and discussion of our evaluation is given (Section IV-G).

A. DATASETS

1) PRODUCT DESCRIPTIONS DATA

The product descriptions of 1778 new toothpaste products are retrieved from the Mintel GNPD from 01/01/2012 to 31/12/2020. These descriptions are filtered from an original larger set of product descriptions to only include products available in USA, U.K., Canada and Australia. These nations are selected as Reddit is largely comprised of users from these areas, thus resulting in a fairer experiment.[11]

2) SOCIAL MEDIA DATA

The Reddit social media data is obtained from the Pushshift API [41], where all posts from 01/01/2012 to 31/12/2017 which contained the term "toothpaste" are chosen (*Social Media Target Keyphrase* - Table 2). After all non-English posts are removed from the dataset, its total size is 231,291 posts coming from 170,065 unique users across 8,303 different subreddits. In our experiments, we only mine over the sentences in which the term "toothpaste" is mentioned. We do this as posts on Reddit can be quite large and thus move away from the discussion of toothpaste within a certain window of sentences where the keyphrase is mentioned.

3) GOOGLE TRENDS

In our implementation, in the context of toothpaste needs, the *Google Trends Category* (Table 2) set is "Oral & Dental Care". As with the Reddit data, this data was collected from 01/01/2012 to 31/12/2017.

TABLE 4. Distribution of POS tags for customer needs using spaCy.

Nouns	Adjectives	Verbs	Adverbs
0.557	0.214	0.203	0.026

B. ILLUSTRATIVE EXAMPLE OF METHODOLOGY: TOOTHPASTE CUSTOMER NEEDS

In this section, we describe the steps of the methodology for finding toothpaste customer needs for the Product Descriptions Algorithm (initially described in Section III-B) and the Social Media Algorithm (initially described in Section III-C). In our experiment, the product descriptions algorithm produces lists of customer needs from 2015-02-28 to 2020-12-31 (71 months) while the social media algorithm produces needs from 2015-01-31 to 2017-12-31 (36 months). Both of the algorithms operate on a *Fixed Time Window Length* (Table 2) of 1 month where ranked lists of customer needs are produced each month for both datasets. Both of the algorithm also use 36 months (i.e. 3 years) for the *Number of Past Time Windows* (Table 2) in order to produce needs at each window. The overall task is to determine the level to which the social media algorithm can predict needs at each window 1 - 36 months (*Future Prediction Time* - Table 2) in the future for the product descriptions algorithm, which is the reason why the product descriptions algorithm produces needs 36 months in the future, respective to the time frame the social media algorithm is operating on.

1) PRODUCT DESCRIPTIONS ALGORITHM

As the product descriptions algorithm uses a supervised NER algorithm, we require labelled data with entities expressing customer needs. In order to obtain this labelled data, an annotator (one of the authors) tags customer needs from 90 product descriptions which are used to tune the NER model. As described when detailing the data annotation in Section III-B (Annotation), annotators label "direct" and "indirect" customer needs. In the context of toothpaste products, the annotators seemed to mainly label the direct needs into health problems (bleeding, plaque, gingivitis), health claims (whitening, strengthening, fresh) and other non-health related benefits (recyclable, vegan, kosher) while they mainly labelled indirect needs into flavours (mint, berry, cinnamon) and ingredients (charcoal, fluoride, SLS). The distribution of the main POS tags relating to customer needs is shown in Table 4 (found using the python library spaCy). Low occurring POS tags are not recorded as they are simply used to separate tokens in a tagged entity (e.g. a hyphen with the POS of punctuation in the entity "tea-tree oil"). It is noteworthy that almost all of the tokens are nouns, adjectives, verbs and adverbs. This information is used by the social media algorithm when choosing which POS tags to include/exclude as customer needs (in order to result in a fair experiment).

Similarly to [102], a random sample of 20 of these descriptions are doubly annotated by a second annotator (independent from the study) in order to verify the gold standard by finding the Inter Annotator Agreement (IAA). In our

[10] https://www.statista.com/statistics/326389/global-oral-care-market-size/ - last accessed 01/04/2022

[11] https://www.similarweb.com/website/reddit.com/#overview - last accessed 01/04/2022

study, the second annotator has experience with marketing toothpaste products which is specifically useful when labelling customer needs from these products. Although Cohen's Kappa [103] is considered the standard agreement measure for IAA, it is not the most relevant measure for named entity annotation, as pointed out in many previous studies [104]–[107]. This is because Cohen's Kappa requires the number of negative instances which is unknown for named entities as they are sequences of tokens (no known number of items to consider) [106], [107]. As a solution to this, Cohen's Kappa has been used on the token-level which is known to have two drawbacks: 1) annotators look at sequences of tokens for entities not individual tokens; and 2) the number of "negative" unannotated tokens will be much larger than the "positive" annotated tokens which would overestimate the Kappa score [106], [107]. Due to this, the F1-score (which does not require the number of "negative" unannotated tokens) has been used alongside the token-level kappa [106], [107]. Table 5 shows the token-level F1 and Cohen's Kappa scores which are calculated for the agreement between the two sets of annotations. The token-wise scores are similar to that of other accepted NER datasets in the literature [106], [107]. The token-level kappa indicates "almost perfect" agreement [108] (i.e. the score is in the range of 0.81-1), although the token-wise version is overestimated in this case.

After the data is labelled, it is augmented using all of the techniques previously discussed in the data augmentation section in Section III-B (Data Augmentation). The hyperparameters that are used to tune the augmentation methods are *p* (probability of replacing a token) to 0.4 and the *number_of_additional_generated_instances* to 2. These values are chosen as they represent the average or most common values used in the grid search experiments of [71]. The augmented version of the data along with the original training data (810 samples in total) is then used to train each base spaCy model in the ensemble (Run Ensemble of Supervised Models). Each base model is trained for 5 epochs. The ensemble then performs inference over a larger set of product descriptions for each fixed time window of data. After the entities are classified and grouped by date (Run Ensemble of Supervised Model), the product descriptions algorithm applies a minimum document frequency threshold (*Pdesc Min Document Frequency* - Table 2). We let this parameter value be equal to 5 in our experiments. We set this value because there is on average ≈ 593 products over each of the previous 36 month sliding time windows of data. Thus, if a need didn't occur in at least 5 of these 593 products it would be difficult to call it "trending".

2) SOCIAL MEDIA ALGORITHM

The parameter values for the social media algorithm are mainly tuned based on commonly accepted values applied previously in the literature. The remaining values are found based on an analysis from a grid search experiment (i.e. *% Most Similar to Gold Standard Subreddit*, *Social Media Min Document Frequency* and *Min Chi Square P-value*).

TABLE 5. Token-wise F1 and Cohen's Kappa annotator agreement scores.

Token-level F1	Token-level Kappa
0.856	0.824

The algorithm uses the processing library spaCy to tokenize, lemmatize and POS tag words. For the removal of posts based on irrelevant subreddit (Removal of Posts Based on Irrelevant Subreddit), the *Gold Standard Subreddit* (Table 2) used in the context of finding toothpaste customer needs is r/Dentistry.[12] This subreddit is chosen as it is most likely to discuss "toothpaste" customer needs. In order to remove bots and moderators (Removal of Posts based on Author), simple regex rules are used in order to detect authors with the words "mod" or "bot" in their usernames. However, more complex ways of detecting these post authors have been proposed in the literature (e.g. bot detection [109]). The "Phrase Detection" model in Gensim is used with the default parameters for both the minimum NPMI score and minimum frequency in order to collocate words together (Collocation Detection).[13] In the literature, it is common to restrict keyphrases which relate to product features (indirect customer needs) to only noun phrases [110]–[112]. However, since our ground truth contains nouns, adjectives, verbs and adverbs (as shown in Table 4), we apply the flexible restriction that any phrase can be a customer need as long as it contains words with these POS tags, thus allowing our *Allowed POS Tags* (Table 2) parameter to be any of these POS tags (Linguistic Pattern Matching). Stop words, URLs and curse words are all removed by the algorithm (Non Customer Need Word Removal). As in [113], phrases which are stop words, according to spaCy's list of stop words are removed. URLs are removed using regex rules. Finally, curse or profanity words are removed with the better-profanity library which uses a simple word list to detect profane words.[14] In our approach to finding non-discriminative phrases (Removal of Non-Discriminate Phrases), we use the python library wordfreq (representative of a normal distribution of words) to act as our external reference corpus.[15]

For the remaining parameters we perform an exhaustive grid search, in order to determine what combinations of hyper-parameter values resulted in the best ranking of keyphrases for the social media algorithm i.e. that correspond with top keyphrases in future product descriptions. The hyper-parameter values we use for each parameter in the experiment are recorded in Table 6. The table shows the minimum value, the maximum value and the step size from the minimum to the maximum values we use for the remaining values (e.g. the *Min Chi Square P-value* parameter

[12] https://www.reddit.com/r/Dentistry/ - last accessed 01/04/2022

[13] https://radimrehurek.com/gensim/models/phrases.html - last accessed 01/04/2022

[14] https://github.com/snguyenthanh/better_profanity – last accessed 01/04/2022

[15] https://pypi.org/project/wordfreq – last accessed 01/04/2022

uses the values 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1). We determined that the values we use for the *% Most Similar to Gold Standard Subreddit* and the *Social Media Min Document Frequency* are sufficient enough to find reasonable results (100 and 20 values are used for each parameter retrospectively). For the *Min Chi Square P-value*, we search between these values as the values of 0.01, 0.05 and 0.1 are all commonly accepted significance values used in the literature [114]. As this grid search is performed on the test set we understand that this can result in over optimized performance which would not necessarily generalise with the same level of performance. To mitigate any such concerns, we show how a sequential range of values for each parameter yields similar results with little deviation, to demonstrate that it is not a particular set of parameter values that achieve a high performance. In our experiments, we found the parameter value range combinations recorded in the ''Range Values'' column of Table 6 to show good performance. The results produced from these parameter ranges will be detailed later in this section (after the experimental setup is explained).

C. PRODUCT DESCRIPTION ENTITY EXTRACTION EVALUATION

We evaluate the needs produced by the product descriptions algorithm (discussed in Section III-B) in order to show that they are sufficiently rich for the assessment of the social media algorithm (i.e. accurate ground truth). Here, we evaluate the supervised NER customer need's extraction model and not the entire product descriptions algorithm. This is done as it can be assumed that if needs are detected by the supervised algorithm with good precision and recall, then representative needs can be produced by the algorithm at each fixed time window. This is because the ''Infrequent Phrase Removal'' and ''Phrase Importance Ranking'' phases of the product descriptions algorithm are only simple statistical measures.

We evaluate the ensemble of supervised models (composed of 19 models) using 3-fold cross-validation on the annotated data. On each iteration of cross-validation, we run each model in the ensemble for 5 epochs (as done in Section IV-B). As discussed, the training data is augmented for each model using the same parameters as discussed in Section IV-B. In order to evaluate the entities produced by the algorithm with the gold-standard test set, the same ''strict'' NER evaluation strategy as in [115] is used (i.e. exact-boundary and type matching).[16] The precision and recall scores of the strict evaluation with 3-fold cross-validation is 91.64% and 95.35% retrospectively (rounded to 2 decimal places). These results show that the product descriptions algorithm is capable of extracting entities which are representative of customer needs with high accuracy. This proves the needs produced by the algorithm are rich enough for the evaluation of the social media algorithm, thus demonstrating that the results from the social media algorithm are valid in further experiments.

D. EXPERIMENTAL SET-UP AND RESULTS

The aim of the experiment is to evaluate the hypothesis **H** that customer needs produced by the social media algorithm are early indicators of needs as expressed in future product descriptions. As stated in Section IV-B, in the context of toothpaste products, these needs consist of ingredients, flavours, health problems, health claims or other non-related health benefits.

In order to evaluate the hypothesis, we follow a similar methodology to [56], [116], who evaluate their topic detection algorithms by submitting lists of topics in fixed time windows. These approaches are evaluated by comparing their algorithms submitted topics to ground truth topics each time window, therefore allowing metrics like precision and recall to be calculated. Similarly, our approach submits lists of keyphrases each time window and compares them to lists of keyphrases produced by the product descriptions algorithm (ground truth). Other than our approach being different to [56], [116] who evaluate topics (lists of keywords) rather than keyphrases, our approach is also different when comparing the time windows in which to compare an algorithm's output to ground truth labels. [56], [116] make comparisons in the same time window for the purposes of detecting events (e.g. comparing the algorithm's topics and ground truth topics both in March 2016). However, in our approach the social media algorithm's output is compared to outputs produced by the product descriptions algorithm (ground truth) at a future date. This is because the goal of the social media algorithm is to observe whether it can predict customer needs occurring in future product descriptions. Specifically, instead of a customer need being considered matched if it occurs in the same time window produced by the product descriptions, it is considered a match if it occurs 1-36 months (*Future Prediction Time* - Table 2) before the product descriptions. When specifically matching customer needs, similarly to [56], [116], keyphrases are considered a match if they were within a Levenshtein similarity of 0.8. This is to allow for some potential misspelling which can occur on social media.

As it is possible to determine if the keyphrases extracted from both of the algorithms match, standard ML metrics are able to be calculated (e.g. precision, recall etc.) Similarly to [56], [116], our metrics are calculated over reduced numbers of produced keyphrases by the social media and the product descriptions algorithms. This is due to the fact that the keyphrases produced by each of the algorithms are ranked. Specifically in our experiment, the number of keyphrases K we use to match from each of the algorithms is 5, 10, 15 and 20. We feel that this is a large enough range in order to find needs which are highly important (e.g. top 5 needs) and also needs which are slightly less important but still relevant (e.g. top 20 needs).

In our experiments, we record two main metrics a) mean precision and b) recall. In order to calculate mean precision, we first calculate precision. Precision is calculated at each fixed time window (i.e. each month) and is defined as the

[16] https://github.com/davidsbatista/NER-Evaluation – last accessed 01/04/2022

TABLE 6. Grid search hyper-parameter values.

Parameter Name	Min Value	Max Value	Step Size	Range Values
% Most Similar to Gold Standard Subreddit	0.01	1.0	0.01	0.05-0.20
Social Media Min Document Frequency	0.00005	0.00025	0.00001	0.00005-0.00020
Min Chi Square P-value	0.01	0.1	0.01	0.01-0.03

This exhaustive grid search searches over each parameter for the values ranging from the "*Min Value*" column to the "*Max Value*" column with a step size specified by the "*Step Size*" column. The value ranges for each parameter that we use in our experiments are contained in the "*Range Values*" column.

number of correct keyphrases the social media algorithm is able to find (1-36 months ahead of the date in which the keyphrases are found) divided by the number of keyphrases K it produced. As precision is calculated on a monthly basis, mean precision over the 3 years (2015-2017) is able to be computed, which is defined as the average of the precision scores. In order to compute recall, we define it as the total number of unique keyphrases the social media algorithm is able to match in future product descriptions (i.e. 1-36 months ahead) divided by the total unique number of keyphrases K the product descriptions algorithm produced in the entirety of the analysis. For this metric it is important to reiterate that the social media algorithm produced keyphrase needs from 2015-2017 while the product descriptions algorithm produced keyphrase needs from 2015-2020. It is not feasible to compute recall over multiple fixed time windows, thus computing "mean recall" (as with mean precision). This is because the number of unique terms 1-36 months ahead for the product descriptions algorithm is far greater than the K unique terms produced by the social media algorithm at a particular fixed time window.

The experiment proposes two approaches for finding keyphrases on social media, our approach (A) and a baseline approach (B). As shown in Table 6, our approach uses a range of consecutive values for three different parameters. This is done to show that a wide range of parameter values can be used for our approach and still achieve good performance rather than any one set of parameter values (as discussed in Section IV-B). For the baseline approach, each month the algorithm produces the most frequent lemmatized unigrams and bigrams which occur in that month. As in Section IV-B, the same restriction that the chosen unigrams and bigrams can only contain nouns, adjectives, verbs and adjectives is followed. This is carried out as the entities tagged from the product descriptions only contain these POS tags (as shown in Table 4) and thus results in a fair experimental comparison. We use this baseline approach to provide context (in the absence of any other available baseline in the literature) for our precision and recall results.

This baseline approach is also used in order to illustrate how difficult it is to find future customer needs and thus achieve any non-zero results in our defined metrics. The task itself is very challenging as users rarely discuss needs relating to products on social media. The task thus really tests our algorithms performance at finding customer keyphrase needs in the noise of many more irrelevant keyphrases i.e. a needle in a haystack problem. What further compounds the difficulty of the task is the way in which our evaluation must be carried out in order to match keyphrase needs (i.e. match keyphrase strings within a Levenshtein distance of 0.8). Moreover, users on social media sometimes don't use the same vocabulary for a customer need as is detected in product descriptions e.g. the product descriptions algorithm detecting the toothpaste ingredient "Sodium Lauryl Sulfate" which might correspond to the abbreviated name of "SLS" detected by the social media algorithm. Therefore, keyphrase needs can go undetected if they carry a similar meaning however aren't spelt similarly. The specific way in which we also define our recall metric along with the way in which our social media algorithm works makes it inherently difficult to achieve high results. Recall only gets increased results for finding future keyphrases at the time which haven't been detected before by the social media algorithm, thus preferring a diverse set of needs to be produced. However, the social media algorithm produces highly similar keyphrases between adjacent fixed time windows. This is because they all work on data from previous time windows (36 months in our experiments) meaning adjacent time windows work on highly similar data subsets. Therefore, the keyphrases they produce are similar. Furthermore, the fact that the social media algorithm only produces customer needs from 2015-2017 while the product descriptions algorithms produces needs from 2015-2020 further increases the difficulty of the metric i.e. social media algorithm must detect keyphrase needs far into the future.

Table 7 shows the mean precision and recall results for our approach (A) and the baseline approach (B) over increasing values of K (rounded to 3 decimal places). As our approach is recorded over a range of different parameter values, the table shows the mean results from these parameter values. Thus, the results we present are not stylised (or best case) results, but rather give a general trend or indication of performance. The mean precision results show that our approach is considerably better than the baseline approach, which is to be expected. However, the mean precision in which needs are identified across the parameter values is impressive, with future needs being found with a mean result between 10-15% across all values of K. The fact that mean precision for our

TABLE 7. Mean precision and recall results for our approach (A) vs Baseline (B).

	Mean Precision				Recall			
	$K=5$	$K=10$	$K=15$	$K=20$	$K=5$	$K=10$	$K=15$	$K=20$
A	0.158	0.141	0.118	0.103	0.021	0.024	0.034	0.046
B	0.000	0.000	0.000	0.026	0.000	0.000	0.000	0.008

Our Approach (A) shows the mean performance when searched within the parameter values in the "Range Values" Column of Table 6 (768 records). The Baseline Approach (B) shows the performance when the most frequent lemmatized unigrams and bigrams which occur in a given months are produced.

TABLE 8. Evaluation of social media algorithm on top customer needs ($K=20$).

Need	Mean Num. Times Detected	Mean 1st Date Social Media Detected	1st Date Product Description Detected
charcoal	25.238	2015-01-31	2017-10-31
coconut	16.357	2015-03-31	2017-08-31
enzyme	0	n/a	2018-07-31
bamboo	0.151	2017-08-31	2017-11-30
eco-friendly	1.368	2015-10-31	2015-02-28
vegan	22.103	2016-02-29	2015-04-30

Columns "Mean Num Times Detected" (0-36 months) and "Mean 1st Date Social Media Detected" (2015-01-31 - 2017-12-31) show the mean performance when searched within the parameter values in the "Range Values" Column of Table 6 (768 records)

approach gets better for lower values of K shows that there is success in the ranking of these keyphrases i.e. on average the matched keyphrases appear near the top of the produced lists. The recall results for our approach are also better than the baseline across all values of K. A possible reason the recall results are not as impressive (compared to the precision results) may be due to the fact that similar results are often produced across adjacent time windows when the Mann-Kendall slope approach is used (Phrase Importance Ranking). This factor may contribute to the low recall results in general. That being said, the way in which recall is defined in our experiments makes it very difficult to achieve high results in this metric. As the results in Table 7 only show the mean results across a subset of parameter values, we go on to show the distribution of these results across the chosen parameter values in order to show that the results don't deviate much from each other (see V).

E. CASE STUDY EVALUATION: TOOTHPASTE CUSTOMER NEEDS

Prior to carrying out any experimentation, a product discovery team from a large undisclosed Multinational Corporation (MNC) specializing in the oral-care sector provided us with a list of the top customer needs from the time period 2015-2020. These needs along with selected keyphrases associated with them (contained in brackets) are: 1) charcoal toothpaste (charcoal, activated charcoal); 2) coconut toothpaste (coconut, coconut oil); 3) enzyme-boosted products (enzyme, enzymatic); 4) bamboo toothbrushes (bamboo); 5) eco-friendly products (eco, eco-cleaner, eco-friendly, biodegradable, sustainable, recyclable, natural, waste, plastic); and 6) vegan-based products (vegan). We carry a retrospective analysis out to observe whether these needs can be detected by the social media algorithm earlier than the product descriptions algorithm. Here, a need from list is considered to be found if any of its associated keyphrases match the keyphrases produced by the social media algorithm each month e.g. the need "enzyme-boosted" is considered to be found if a keyphrase from the algorithm contains either "enzyme" or "enzymatic".

For an analysis which we carry out between 2015-2017 (36 months), we test the social media algorithm to observe whether it can detect the needs in the top needs list within the hyper-parameter value ranges (i.e. "Range Values" column) recorded in Table 6. Table 8 goes on to show the total mean number of times these needs are detected (Mean Num Times Detected), the mean first date the need is detected by the social media algorithm (Mean 1st Date Social Media Detected) and the first date the need is detected by the product descriptions algorithm (1st Date Product Descriptions Detected) when the number of produced keyphrases K is 20 for both the product description and social media algorithm. As this experiment records over a range of hyper-parameters, the table shows the mean values for the social media algorithm's recorded columns (i.e. "Mean Num Times Detected" and "Mean 1st Date Social Media Detected"). The "Mean Num Times Detected" column is rounded to 3 decimal places.

The precision the algorithm is able to detect the needs is good with 4 out of the 6 needs being detected with an average value of at least 1 time in the analysis (charcoal, coconut, eco-friendly and vegan), while 5 of the needs are detected at least once within the defined hyper-parameter range (charcoal, coconut, eco-friendly, vegan and bamboo). The number of times the algorithm is able to detect some of the needs is also good with charcoal, coconut and vegan being detected on average 25, 16 and 22 times in a total of 36 time windows (i.e. months). Some of the lead times for the needs are impressive with charcoal and coconut having average lead times by 58 and 53 months ahead before they are detected by the product descriptions algorithm. This realization is significant as it would allow companies to identify needs long before they become mainstream, thus giving them several advantages in the marketplace. The bamboo and enzyme needs are rarely or never detected by the social media algorithm. The need of bamboo may not have been found often as it is a need for toothbrush rather than toothpaste products (our analysis searches for toothpaste needs). While in the case of enzyme, this may have been because users simply didn't discuss this need on social media and in search engines, even if it was identified by large corporations through other methods (e.g. user interviews or questionnaires). As the results in Table 8 only show the mean results across a subset of parameter values, we go on to show the distribution of these results across the chosen parameter values in order to show that the results don't deviate much from each other (see V).

F. IMPORTANCE OF STEPS OF THE METHODOLOGY

In this section, reasons for certain steps being taken in the social media algorithm are explained. We do this by detailing why certain methods are used or why special data sources are employed and inspect the impact they have on the task of predicting future customer needs. Specifically, we look at two steps of the methodology: a) Removal of Posts Based on Irrelevant Subreddit(s), and b) Phrase Importance Ranking From Reddit & Google Trends.

1) REMOVAL OF POSTS BASED ON IRRELEVANT SUBREDDIT(s)

A key parameter in the *Removal of Posts Based on Irrelevant Subreddit(s)* step of the social media algorithm is the *% Most Similar to Gold Standard Subreddit* parameter (Table 2). This parameter controls the number of posts to be included/excluded in the remainder of the analysis. In order to show how this step of the methodology impacts the results, we show the results when various different consecutive ranges of values for this parameter are used. These results are recorded in Table 9 (rounded to 3 decimal places), which use the same experimental set-up and metrics as in Table 7. We also show the results when this approach isn't used and thus no posts are removed from the analysis i.e. *% Most Similar to Gold Standard Subreddit*=1. In order to observe how this parameter solely impacts the results, we consider the complete ranges for the remaining two grid searched parameters i.e. *Social Media Min Document Frequency*=0.00005-0.00025 and *Min Chi Square P-value*=0.01-0.1. As each row consists of a range of different parameter values, we record the mean results for each column (as in Table 7).

There seems to generally be an increase in results when this step of the methodology is performed compared to when it isn't. In order to show this, we run a set of significance tests comparing the results when *% Most Similar to Gold Standard Subreddit* is between 0.01-0.99 (i.e. step is performed) compared to when *% Most Similar to Gold Standard Subreddit*=1 (i.e. step isn't performed). Specifically, we compare samples of the mean precision and recall results generated by both approaches over the same values of produced keyphrases *K* used in Table 9 (i.e. 5, 10, 15 and 20). For each result set, we run a Mann-U-Whitney Rank Test. We use this instead of a t-test because the result data is not normally distributed nor contains the same sample sizes. From the set of tests, the p-values comparing each of the results all have a value less than 0.01. After the tests are run, we also find that the mean results of all the samples are all greater in favour of using the approach (i.e. *% Most Similar to Gold Standard Subreddit* between 0.01-0.99). We can therefore conclude that using the data reduction approach provides a positive statistical difference in performance. This parameter can thus be significant in the process of finding future trends and is thus a contribution of our work.

2) PHRASE IMPORTANCE RANKING (REDDIT & GOOGLE TRENDS)

In this section we verify the method of keyphrase ranking for the social media algorithm using data from both Reddit and Google Trends (as discussed in Section III-C). In our approach, we perform data reduction and keyphrase extraction using data from only Reddit and then after consider the slope values from both Reddit and Google Trends for a keyphrase in order to rank it in comparison with other keyphrases. Here we detail our reason for taking this keyphrase ranking approach while at the same time proving that either one of the data sources on their own can effectively rank the keyphrases and obtain reasonable results. We do this by showing the performance of keyphrase ranking using the slope values from a) both data sources as in our methodology i.e. *Reddit + Both Ranking*; b) only Reddit i.e. *Reddit + Reddit Ranking*; and c) only Google Trends i.e. *Reddit + Google Trends Ranking*. For each of these ranking methods, the frequency based time series associated with each keyphrase is normalized using unit vector normalization (as in described Section III-C). This is done because these time series are based on raw keyphrase frequencies, meaning mainly highly frequent keyphrases appear nearer to the top of the final output lists for each fixed time window if no normalization occurs. This happens as the method of keyphrase ranking ranks the keyphrases based on the slope value returned from the MK trend test for their retrospective time series. If the time series are between 0-1 (normalized) then keyphrases which show the most relative increase in keyphrase usage are picked. However, if the time series is between 0-keyphrase frequency (no normalization) then keyphrases which show an overall increase in usage are picked (mostly highly frequent time series). We use the entire range of hyper-parameters recorded in the grid search search experiment (Table 6) rather than the hyper-parameter value ranges used in the experiment (''Range Values'' column in Table 6). We do this as the hyper-parameter value ranges used in the experiment are optimized to work on the *''Reddit + Both Ranking''* method. For each of these methods across a range of different hyper-parameter value ranges we record the mean performance of the mean precision and recall scores across the same range of number of produced keyphrases *K* as in Table 7. These results are recorded in Table 10 (rounded to 3 decimal places).

TABLE 9. Mean performance of the mean precision and recall results over multiple values of % *most similar to gold standard subreddit* (%MSGSS) parameter.

% MSGSS	Mean Precision				Recall			
	$K = 5$	$K = 10$	$K = 15$	$K = 20$	$K = 5$	$K = 10$	$K = 15$	$K = 20$
0.01 - 0.1	0.148	0.135	0.114	0.104	0.017	0.025	0.040	0.054
0.11 - 0.2	0.161	0.139	0.117	0.102	0.023	0.024	0.034	0.042
0.21 - 0.3	0.149	0.132	0.112	0.101	0.020	0.024	0.030	0.044
0.31 - 0.4	0.160	0.140	0.112	0.104	0.020	0.023	0.029	0.048
0.41 - 0.5	0.150	0.135	0.119	0.104	0.019	0.021	0.033	0.044
0.51 - 0.6	0.147	0.129	0.112	0.102	0.017	0.020	0.029	0.038
0.61 - 0.7	0.148	0.135	0.120	0.109	0.016	0.020	0.029	0.039
0.71 - 0.8	0.151	0.136	0.117	0.106	0.016	0.018	0.028	0.039
0.81 - 0.9	0.146	0.131	0.114	0.101	0.016	0.016	0.028	0.037
0.91 - 0.99	0.145	0.127	0.112	0.100	0.015	0.016	0.028	0.038
1	0.144	0.127	0.112	0.100	0.015	0.016	0.028	0.039

Results show mean performance when searched within complete ranges of the remaining two parameter values in Table 6 i.e. *Social Media Min Document Frequency*=0.00005-0.00025 and *Min Chi Square P-value*=0.01-0.1. There mean for each row is calculated based upon 2600 records.

TABLE 10. Mean performance of the mean precision and recall results for ranking methods.

	Mean Precision				Recall			
	$K = 5$	$K = 10$	$K = 15$	$K = 20$	$K = 5$	$K = 10$	$K = 15$	$K = 20$
Reddit + Both Ranking	0.150	0.134	0.115	0.103	0.018	0.021	0.031	0.042
Reddit + Reddit Ranking	0.088	0.102	0.103	0.100	0.028	0.039	0.058	0.073
Reddit + Google Trends Ranking	0.114	0.104	0.089	0.086	0.010	0.020	0.028	0.030

Results show mean performance when searched within complete ranges of the parameter values in Table 6 i.e. *% Most Similar toGold Standard Subreddit*=0.01-1, *Social Media Min Document Frequency*=0.00005-0.00025 and *Min Chi Square P-value*=0.01-0.1. The mean for each row in calculated based upon 26000 records.

The Reddit ranking approach performs better than the Google Trends ranking for the recall metric due to the way in which it is calculated. In our experiments recall is calculated over the entire time period in which the analysis is done, in which the social media algorithm tries to detect every phrase ever produced by the product descriptions algorithm. This thus would benefit from a larger unique number of phrases being produced by the social media algorithm, as there is then a higher chance of it being found as a keyphrase produced by the product descriptions algorithm. It is the volatility in the keyphrase frequencies of the Reddit data leads to the Reddit ranking approach performing better than the Google Trends ranking approach, obtaining better results in the recall metric across every value of K. Google Trends ranking on the other hand leads to a better performance compared to the Reddit ranking for some values of K for mean precision. This could be due to the fact that Google Trends is made up of more people than Reddit and thus the slope value returned for its ranking could be more precise for future needs prediction. The reason why this analysis uses the "Both" method of ranking is due to it being a good trade-off between the mean precision and recall scores produced by the Reddit and Google Trends ranking methods. The "Both" ranking method gets better mean precision on every iteration compared to the Google Trends ranking method and on some iterations it is close to the recall results produced by the Reddit data.

G. SUMMARY AND DISCUSSION

To help guide our evaluation we proposed two research questions: 1) can future customer needs (as identified in future product descriptions) be predicted using UGC from social media; and 2) is our defined data reduction approach useful for the task of predicting future customer needs (described in Section III-C).

To address these questions, we first tailored our approach (datasets and the implementation of the methodology) specific to "toothpaste" customer needs (Section IV-A and IV-B). Using product description data from GNPD, we performance tested our approach for entity extraction from product descriptions in order to show that customer needs could be extracted with high precision and recall from product descriptions (Section IV-C). This is key as it illustrates that these needs represent an appropriate ground truth (of future needs as expressed in new product trends) for the evaluation of the social media algorithm. Whilst we recognise that there may be other latent causal relationships (such as other forms of investigative product development resulting in new products) this approach enables us to illustrate that trends on social media are (at least) linked to future products.

Using timestamped new product needs data (i.e. needs expressed via product description keywords over time) as a ground truth of customer needs, we explored the ability of our social media algorithm to predict customer needs (i.e. keywords detected from UGC on Reddit) between 1 and 36 months prior to their appearance in new products. Our results illustrate that customer needs could be extracted with precision and recall scores significantly better then our defined baselines (Section IV-D). To assist in the validation of these findings, we approached a large MNC in the oral-care sector to provide a set of the "biggest" historical new product trends in the market. Taking these as a case study, we showed that our approach was capable of finding such ''high-impact'' needs often with large lead times. Thus, we would argue that our approach can indeed find future customer needs using social media data (question 1).

It would, however, be easy to argue that with such large amounts of data, anything can be correlated with a set of keywords in product descriptions. We note a few details of our approach that specifically attempt to hinder such eventualities. Firstly, our approach uses ranked keywords and this ranking is incorporated into the derivation of evaluation metrics. Secondly, our performance metrics are quite restrictive: it is difficult to score ''high''. Thirdly, we explore a variety of hyperparameter settings and illustrate that our approach is not sensitive to optimally picked values for these parameters. Finally, our approach to data reduction seeks to eject arbitrary (but still potentially relevant) data from consideration (question 2). We have shown this as well as other key steps on our approach (Section IV-F) to result in an effective methodology for finding future customer needs using social media data (Reddit).

V. CONCLUSION

Past approaches mine current customer needs from static collections of UGC. Because of this, there is a lack of research on analyzing customers' needs over time and predicting their future needs. Therefore, in this paper, we outline an approach for predicting future customer needs on Reddit and analyzing them over time. We do this by framing the problem of extracting customer needs as a keyphrase extraction problem where future ranked lists of keyphrases are produced at fixed time windows. Needs were found for a given product type by only considering posts which contained the presence of a user-defined keyword and were also included in particular subreddits which were likely to discuss topics relating to the needs of the product type. In order to extract keyphrases from UGC, various techniques from keyphrase extraction were proposed. These keyphrases were then ranked using knowledge from UGC (Reddit and Google Trends) and applying a Mann-Kendall trend test in order to discover whether a keyphrase's frequency was increasing over time. A separate customer needs extraction algorithm was also run over a dataset of new-to-market product descriptions in order to observe whether there was a future time lagged relationship between these needs and the ones occurring on social media. In order to evaluate our approach the domain of ''toothpaste needs'' was studied. It was shown that social media could detect needs occurring in future toothpaste product descriptions with significantly better precision and recall results than our defined baselines. Furthermore, social media was able to detect 4 out of the 6 needs given to us by a large MNC with a mean occurrence of at least once within the hyper-parameter range defined in the experiment. In addition, it was able to detect 2 of these needs with significant lead times over the product descriptions algorithm. The impact of this work is that we evidence that discussions on social media may give companies significant margins to outpace their competitors in new product development. Even doing so by a number of weeks, can be significant in terms of potential profit and market acquisition.

The contributions of this research are as follows:

- Analyzing needs over time - where previous approaches mined needs from static collections of UGC
- Predicting the future importance of needs - with prior studies only detecting the current needs of users
- A novel data reduction approach - which was proven to significantly increase the performance of future need detection
- Proposing a novel evaluation strategy for the prediction of future customer needs by extracting ground truth needs from product descriptions

This research also proposes a framework which allows product development teams to identify future business ideas, therefore resulting in opportunity identification contributions of the research as well.

In light of these contributions, there were also various limitations, indicating areas of future work. In our evaluation, we applied our methodology to the oral-care sector only. Thus, an obvious avenue for future work is to explore other product types (e.g. smartphones). As the techniques used in this approach were based on general data analysis it would be assumed that it would have all the right indicators of generalizing to new product types, however as experiments were not conducted, it is a limitation of the study. Future iterations of this work thus intend to explore this in order to bring more validity to the approach. Also, with many previous studies employing some knowledge of sentiment when extracting customer needs [12]–[17], [19]–[25], [30], [34], [44], [45], [53], [63], [64], it would be interesting to observe whether it is a predictive factor in discovering needs in future product descriptions. Finally, this study has a limitation in the fact that customer needs are being predicted retrospectively. Careful consideration is taken in the treatment of future data as a ground truth when performing our analysis that future information does not bias the analysis. The results materializing from this experiment could thus have similar criticisms to that of predicting election results post facto [117]. That being said, our analysis is of interest nonetheless as we were able to correlate past customer needs on social media to future needs in product descriptions (which to the best of our knowledge hasn't been done before).

APPENDIX A

The results for the parameter range values in Table 6 (i.e. Range Values Column) don't deviate much from the mean. In order to illustrate this, in Figure 3, we plot the distribution of the results for this parameter range for the mean results in Table 7 (i.e. approach A). The figure shows multiple distributions of results for the mean precision and recall values over the various values of K, as recorded in the initial experiment (i.e. 5, 10, 15, and 20). Above each distribution, the maximum

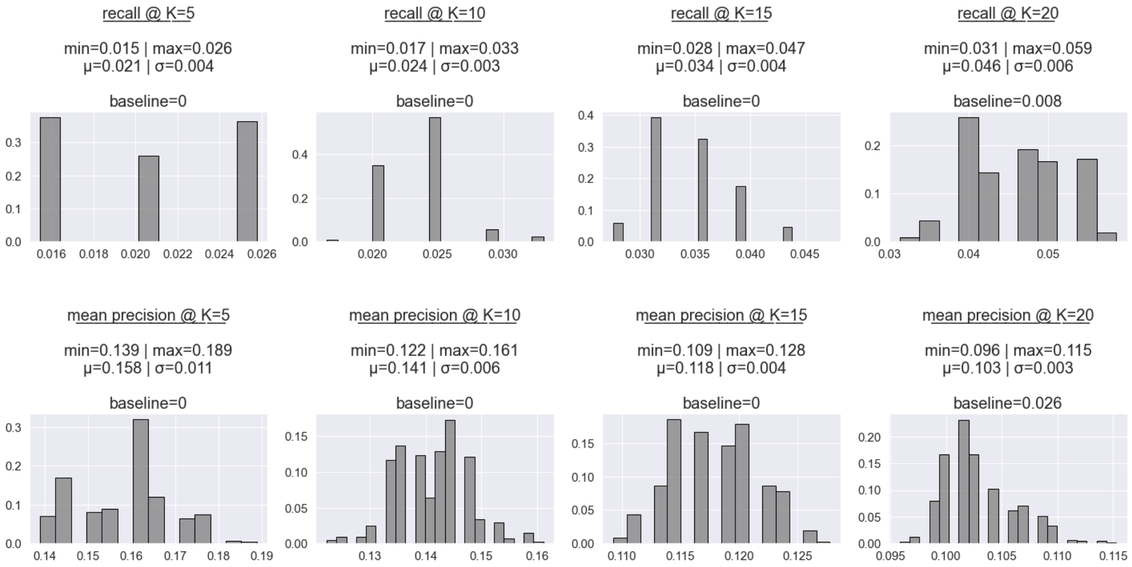

FIGURE 3. Table 7 results distribution.

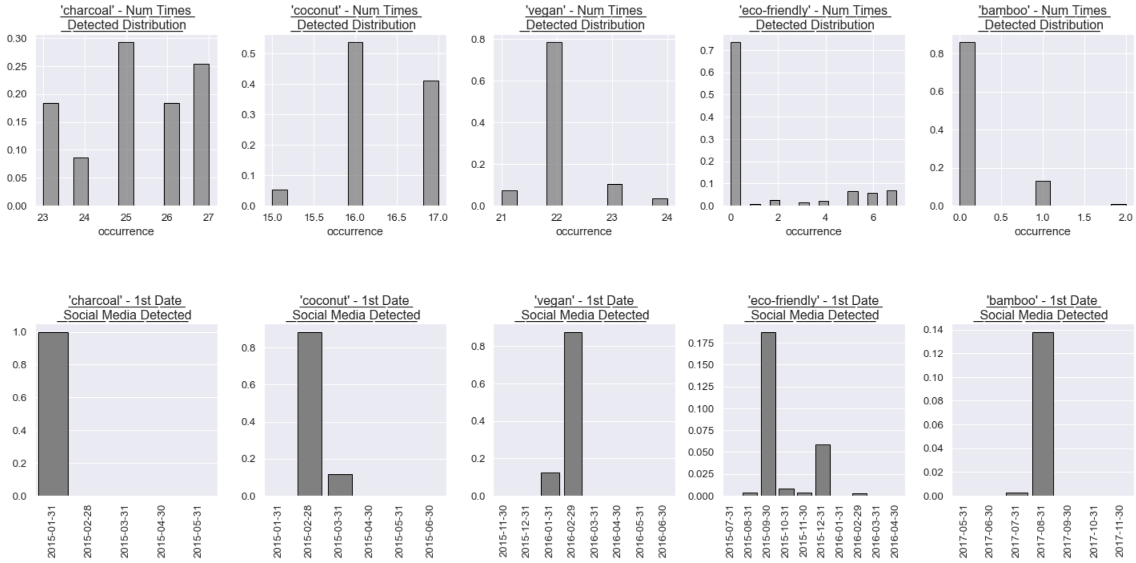

FIGURE 4. Table 8 results distribution.

(max), minimum (min), mean (μ) and the standard deviation (σ) are shown for the results. We also provide the result of the baseline above each distribution (i.e. approach B as recorded in Table 7). We don't show the baseline result as a distribution as it is a single value. Along with the fact that the results don't seem to deviate too much from each other it's also noteworthy that the minimum results for each metric of the algorithm over every value of K performs better than the baseline.

APPENDIX B

The results for the parameter range values in Table 6 (i.e. Range Values Column) don't deviate much from the mean. In order to illustrate this, in Figure 4, we plot the distribution of the results of the "Mean Num Times Detected" and "Mean 1st Date Social Media Detected" columns in Table 8. The figure shows multiple distributions only for the for the needs which were detected by the algorithm i.e. "charcoal", "coconut", "vegan", "eco-friendly" and "bamboo". The results can deviate slightly in some cases (e.g. the "1st Date Social Media" column for the need "eco-friendly"), however, the results generally show low deviation.

REFERENCES

[1] R. Tagiuri and J. A. Davis, "On the goals of successful family companies," *Family Bus. Rev.*, vol. 5, no. 1, pp. 43–62, Mar. 1992.

[2] Y. P. Freund, "Critical success factors," *Planning Rev.*, vol. 16, pp. 20–23, Apr. 1988.

[3] J. Munch, S. Trieflinger, and B. Heisler, "Product discovery–building the right things: Insights from a grey literature review," in *Proc. IEEE Int. Conf. Eng., Technol. Innov. (ICE/ITMC)*, Jun. 2020, pp. 1–8.

[4] J. Melanie Joh and M. Mayfield, "The discipline of product discovery: Identifying breakthrough business opportunities," *J. Bus. Strategy*, vol. 30, nos. 2–3, pp. 70–77, Feb. 2009.

[5] B. J. Zirger and M. A. Maidique, ''A model of new product development: An empirical test,'' *Manage. Sci.*, vol. 36, no. 7, pp. 867–883, Jul. 1990.

[6] H. Takeuchi and I. Nonaka, ''The new new product development game,'' *Harvard Bus. Rev.*, vol. 64, no. 1, pp. 137–146, 1986.

[7] W. Chang and S. A. Taylor, ''The effectiveness of customer participation in new product development: A meta-analysis,'' *J. Marketing*, vol. 80, no. 1, pp. 47–64, Jan. 2016.

[8] M. K. Poetz and M. Schreier, ''The value of crowdsourcing: Can users really compete with professionals in generating new product ideas?'' *J. Product Innov. Manage.*, vol. 29, no. 2, pp. 245–256, 2012.

[9] H. Nishikawa, M. Schreier, and S. Ogawa, ''User-generated versus designer-generated products: A performance assessment at muji,'' *Int. J. Res. Marketing*, vol. 30, no. 2, pp. 160–167, Jun. 2013.

[10] A. Kelly, ''Requirements, discovery, and demand,'' in *The Art Agile Product Ownership*. Berkeley, CA, USA: Springer, 2019, pp. 31–38.

[11] W. Liu, S. Ye, and J. Moultrie, ''Exploring traditional and new web-based methods to involve customers in new product development,'' *Int. J. Product Develop.*, vol. 23, no. 1, pp. 42–64, 2019.

[12] D. Chen, D. Zhang, and A. Liu, ''Intelligent Kano classification of product features based on customer reviews,'' *CIRP Ann.*, vol. 68, no. 1, pp. 149–152, 2019.

[13] M.-C. Chiu and K.-Z. Lin, ''Utilizing text mining and kansei engineering to support data-driven design automation at conceptual design stage,'' *Adv. Eng. Informat.*, vol. 38, pp. 826–839, Oct. 2018.

[14] W. Kim, T. Ko, I. Rhiu, and M. H. Yun, ''Mining affective experience for a kansei design study on a recliner,'' *Appl. Ergonom.*, vol. 74, pp. 145–153, Jan. 2019.

[15] F. Zhou and R. J. Jiao, ''Latent customer needs elicitation for big-data analysis of online product reviews,'' in *Proc. IEEE Int. Conf. Ind. Eng. Eng. Manage. (IEEM)*, Dec. 2015, pp. 1850–1854.

[16] K. Jiang and Y. Li, ''Mining customer requirement from online reviews based on multi-aspected sentiment analysis and Kano model,'' in *Proc. 16th Dahe Fortune China Forum Chin. High-educational Manage. Annu. Academic Conf. (DFHMC)*, Dec. 2020, pp. 150–156.

[17] Z.-J. Zha, J. Yu, J. Tang, M. Wang, and T.-S. Chua, ''Product aspect ranking and its applications,'' *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1211–1224, May 2014.

[18] V. R. Hananto, S. Kim, M. Kovacs, U. Serdult, and V. Kryssanov, ''A machine learning approach to analyze fashion styles from large collections of online customer reviews,'' in *Proc. 6th Int. Conf. Bus. Ind. Res. (ICBIR)*, May 2021, pp. 153–158.

[19] J. Joung and H. M. Kim, ''Automated keyword filtering in latent Dirichlet allocation for identifying product attributes from online reviews,'' *J. Mech. Design*, vol. 143, no. 8, pp. 1–6, Aug. 2021.

[20] J. J. C. Aman, J. Smith-Colin, and W. Zhang, ''Listen to E-scooter riders: Mining rider satisfaction factors from app store reviews,'' *Transp. Res. D, Transp. Environ.*, vol. 95, Jun. 2021, Art. no. 102856.

[21] W.-K. Chen, D. Riantama, and L.-S. Chen, ''Using a text mining approach to hear voices of customers from social media toward the fast-food restaurant industry,'' *Sustainability*, vol. 13, no. 1, p. 268, Dec. 2020.

[22] H.-J. Kwon, H.-J. Ban, J.-K. Jun, and H.-S. Kim, ''Topic modeling and sentiment analysis of online review for airlines,'' *Information*, vol. 12, no. 2, p. 78, Feb. 2021.

[23] B. Jeong, J. Yoon, and J.-M. Lee, ''Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis,'' *Int. J. Inf. Manage.*, vol. 48, pp. 280–290, Oct. 2019.

[24] N. Ko, B. Jeong, S. Choi, and J. Yoon, ''Identifying product opportunities using social media mining: Application of topic modeling and chance discovery theory,'' *IEEE Access*, vol. 6, pp. 1680–1693, 2017.

[25] J. Choi, S. Oh, J. Yoon, J.-M. Lee, and B.-Y. Coh, ''Identification of time-evolving product opportunities via social media mining,'' *Technol. Forecasting Social Change*, vol. 156, Jul. 2020, Art. no. 120045.

[26] S. Tuarob and C. S. Tucker, ''Fad or here to stay: Predicting product market adoption and longevity using large scale, social media data,'' in *Proc. Int. Design Eng. Tech. Conf. Comput. Inf. Eng.*, vol. 55867, Feb. 2013, pp. 1–5.

[27] S. Tuarob and C. S. Tucker, ''Quantifying product favorability and extracting notable product features using large scale social media data,'' *J. Comput. Inf. Sci. Eng.*, vol. 15, no. 3, pp. 1–17, Sep. 2015.

[28] T. Ko, I. Rhiu, M. H. Yun, and S. Cho, ''A novel framework for identifying Customers' unmet needs on online social media using context tree,'' *Appl. Sci.*, vol. 10, no. 23, p. 8473, Nov. 2020.

[29] X. Han, R. Li, W. Li, G. Ding, and S. Qin, ''User requirements dynamic elicitation of complex products from social network service,'' in *Proc. 25th Int. Conf. Autom. Comput. (ICAC)*, Sep. 2019, pp. 1–6.

[30] A. A. Olad and O. Fatahi Valilai, ''Using of social media data analytics for applying digital twins in product development,'' in *Proc. IEEE Int. Conf. Ind. Eng. Eng. Manage. (IEEM)*, Dec. 2020, pp. 319–323.

[31] S. P. Gaskin, A. Griffin, J. R. Hauser, G. M. Katz, and R. L. Klein, *Voice of the Customer*. Hoboken, NJ, USA: Wiley, 2010.

[32] N. Kühl, M. Mühlthaler, and M. Goutier, ''Supporting customer-oriented marketing with artificial intelligence: Automatically quantifying customer needs from social media,'' *Electron. Markets*, vol. 30, no. 2, pp. 351–367, 2019.

[33] N. Káhl and G. Satzger, ''Needmining: Designing digital support to elicit needs from social media,'' 2021, *arXiv:2101.06146*.

[34] H. Jiang, C. K. Kwong, and K. L. Yung, ''Predicting future importance of product features based on online customer reviews,'' *J. Mech. Des.*, vol. 139, no. 11, Nov. 2017, Art. no. 111413.

[35] K. Holt, ''User-oriented product innovation–some research findings,'' *Technovation*, vol. 3, no. 3, pp. 199–208, Aug. 1985.

[36] H. Kärkkäinen, P. Piippo, K. Puumalainen, and M. Tuominen, ''Assessment of hidden and future customer needs in Finnish business-to-business companies,'' *R&D Manage.*, vol. 31, no. 4, pp. 391–407, Oct. 2001.

[37] B. Guo, Y. Liu, Y. Ouyang, V. W. Zheng, D. Zhang, and Z. Yu, ''Harnessing the power of the general public for crowdsourced business intelligence: A survey,'' *IEEE Access*, vol. 7, pp. 26606–26630, 2019.

[38] I. C. D. Morais and E. P. Z. Brito-Eliane, ''Productive consumption and marketplace dynamics: A study in the DIY homemade natural beauty products context,'' ANPAD, São Paulo, Brazil, Tech. Rep., 2015.

[39] D. Freelon, ''Computational research in the post-API age,'' *Political Commun.*, vol. 35, no. 4, pp. 665–668, Oct. 2018.

[40] J. Isaak and M. J. Hanna, ''User data privacy: Facebook, Cambridge analytica, and privacy protection,'' *Computer*, vol. 51, no. 8, pp. 56–59, Aug. 2018.

[41] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, ''The pushshift reddit dataset,'' in *Proc. Int. AAAI Conf. Web Social Media*, vol. 14, 2020, pp. 830–839.

[42] X. Cheng, X. Yan, Y. Lan, and J. Guo, ''BTM: Topic modeling over short texts,'' *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 12, pp. 2928–2941, Dec. 2014.

[43] J. Pang, X. Li, H. Xie, and Y. Rao, ''SBTM: Topic modeling over short texts,'' in *Proc. Int. Conf. Database Syst. Adv. Appl.* Dallas, TX, USA: Springer, 2016, pp. 43–56.

[44] W. M. Wang, Z. Li, Z. G. Tian, J. W. Wang, and M. N. Cheng, ''Extracting and summarizing affective features and responses from online product descriptions and reviews: A kansei text mining approach,'' *Eng. Appl. Artif. Intell.*, vol. 73, pp. 149–162, Aug. 2018.

[45] F. Zhou, J. Ayoub, Q. Xu, and X. Jessie Yang, ''A machine learning approach to customer needs analysis for product ecosystems,'' *J. Mech. Design*, vol. 142, no. 1, pp. 1–13, Jan. 2020.

[46] C. Tucker and H. Kim, ''Predicting emerging product design trend by mining publicly available customer review data,'' in *Proc. 18th Int. Conf. Eng. Design*, 2011, pp. 1–10.

[47] H. Choi and H. Varian, ''Predicting the present with Google trends,'' *Econ. Rec.*, vol. 88, no. 1, pp. 2–9, 2012.

[48] F. Wijnhoven and O. Plant, ''Sentiment analysis and Google trends data for predicting car sales,'' AIS Electron. Library, Seoul, South Korea, Tech. Rep., 2017.

[49] H. Yakubu and C. K. Kwong, ''Forecasting the importance of product attributes using online customer reviews and Google trends,'' *Technological Forecasting Social Change*, vol. 171, Oct. 2021, Art. no. 120983.

[50] V. Gupta, D. Varshney, H. Jhamtani, D. Kedia, and S. Karwa, ''Identifying purchase intent from social posts,'' in *Proc. 8th Int. AAAI Conf. Weblogs Social Media*, 2014, pp. 180–186.

[51] J. Wang, G. Cong, X. W. Zhao, and X. Li, ''Mining user intents in Twitter: A semi-supervised approach to inferring intent categories for tweets,'' in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 339–345.

[52] B. Hollerit, M. Kröll, and M. Strohmaier, ''Towards linking buyers and sellers: Detecting commercial intent on Twitter,'' in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 629–632.

[53] J. Koss and S. Bohnet-Joschko, ''Social media mining in drug development decision making: Prioritizing multiple sclerosis patients' unmet medical needs,'' in *Proc. Hawaii Int. Conf. Syst. Sci.*, HI, USA, 2022, doi: 10.24251/HICSS.2022.368.

[54] Y. Ohsawa, "Chance discoveries for making decisions in complex real world," *New Gener. Comput.*, vol. 20, no. 2, pp. 143–163, Jun. 2002.

[55] J. S. Pinegar, *What Customers Want: Using Outcome-Driven Innovation to Create Breakthrough Products and Services*. Oxford, U.K.: Wiley, 2006, doi: 10.1111/j.1540-5885.2006.00217.x.

[56] C. Comito, A. Forestiero, and C. Pizzuti, "Bursty event detection in Twitter streams," *ACM Trans. Knowl. Discovery Data*, vol. 13, no. 4, pp. 1–28, Aug. 2019.

[57] A. Guille and C. Favre, "Event detection, tracking, and visualization in Twitter: A mention-anomaly-based approach," *Social Netw. Anal. Mining*, vol. 5, no. 1, p. 18, Dec. 2015.

[58] A. Guille and C. Favre, "Mention-anomaly-based event detection and tracking in Twitter," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2014, pp. 375–382.

[59] M. Nagamachi, "Kansei engineering: A new ergonomic consumer-oriented technology for product development," *Int. J. Ind. Ergonom.*, vol. 15, no. 1, pp. 3–11, 1995.

[60] S. Lin, T. Shen, and W. Guo, "Evolution and emerging trends of kansei engineering: A visual analysis based on CiteSpace," *IEEE Access*, vol. 9, pp. 111181–111202, 2021.

[61] X. Lai, S. Zhang, N. Mao, J. Liu, and Q. Chen, "Kansei engineering for new energy vehicle exterior design: An internet big data mining approach," *Comput. Ind. Eng.*, vol. 165, Mar. 2022, Art. no. 107913.

[62] N. Kano, S. Nobuhiko, T. Fumio, and T. Shinichi, "Attractive quality and must-be quality," *J. Jpn. Soc. Service Qual. Control*, vol. 14, no. 2, pp. 39–48, Apr. 1984.

[63] J.-W. Bi, Y. Liu, Z. P. Fan, and E. Cambria, "Modelling customer satisfaction from online reviews using ensemble neural network and effect-based Kano model," *Int. J. Prod. Res.*, vol. 57, no. 22, pp. 7068–7088, 2019.

[64] M. Zhao, C.-X. Zhang, Y.-Q. Hu, Z.-S. Xu, and H. Liu, "Modelling consumer satisfaction based on online reviews using the improved Kano model from the perspective of risk attitude and aspiration," *Technol. Econ. Develop. Economy*, vol. 27, no. 3, pp. 550–582, Apr. 2021.

[65] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. (WI-IAT)*, vol. 1, Aug./Sep. 2010, pp. 492–499.

[66] E. Solis, "Mintel global new products database (GNPD)," *J. Bus. Finance Librarianship*, vol. 21, no. 1, pp. 79–82, Jan. 2016.

[67] G. Zheng, S. Mukherjee, X. L. Dong, and F. Li, "OpenTag: Open attribute value extraction from product profiles," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1049–1058.

[68] D. Putthividhya and J. Hu, "Bootstrapped named entity recognition for product attribute extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 1557–1567.

[69] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," 2019, *arXiv:1901.11196*.

[70] S. Sharifirad, B. Jafarpour, and S. Matwin, "Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs," in *Proc. 2nd Workshop Abusive Lang.*, 2018, pp. 107–114.

[71] X. Dai and H. Adel, "An analysis of simple data augmentation for named entity recognition," 2020, *arXiv:2010.11683*.

[72] T. G. Dietterich, "Ensemble learning," *Handbook Brain Theory Neural Netw.*, vol. 2, pp. 110–125, Mar. 2002.

[73] X. Yang, H. Zhang, X. He, J. Bian, and Y. Wu, "Extracting family history of patients from clinical narratives: Exploring an end-to-end solution with deep learning models," *JMIR Med. Informat.*, vol. 8, no. 12, Dec. 2020, Art. no. e22982.

[74] J. Copara, N. Naderi, J. Knafou, P. Ruch, and D. Teodoro, "Named entity recognition in chemical patents using ensemble of contextual language models," 2020, *arXiv:2007.12569*.

[75] A. P. Tafti, S. Fu, A. Khurana, G. M. Mastorakos, K. G. Poole, S. J. Traub, J. A. Yiannias, and H. Liu, "Artificial intelligence to organize patient portal messages: A journey from an ensemble deep learning text classification to rule-based named entity recognition," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 1380–1387.

[76] L. M. Romo-Fernández, V. P. Guerrero-Bote, and F. Moya-Anegón, "Co-word based thematic analysis of renewable energy (1990–2010)," *Scientometrics*, vol. 97, no. 3, pp. 743–765, Dec. 2013.

[77] L. Trinquart and S. Galea, "Mapping epidemiology's past to inform its future: Metaknowledge analysis of epidemiologic topics in leading journals, 1974–2013," *Amer. J. Epidemiol.*, vol. 182, no. 2, pp. 93–104, Jul. 2015.

[78] J. M. Kleinberg, "Bursty and hierarchical structure in streams," *Data Mining Knowl. Discovery*, vol. 7, no. 4, pp. 373–397, 2003.

[79] M. Uchitpe, S. Uddin, and C. Lynn, "Predicting the future of project management research," *Social Behav. Sci.*, vol. 226, pp. 27–34, Jul. 2016.

[80] S. Krishnamoorthy, "Linguistic features for review helpfulness prediction," *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3751–3759, May 2015.

[81] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in *Proc. 4th Int. Conf. Cyber IT Service Manage.*, Apr. 2016, pp. 1–6.

[82] K. E. Anderson, *Ask Anything: What is Reddit*. London, U.K.: Library Hi Tech News, 2015.

[83] R. Gorwa and D. Guilbeault, "Unpacking the social media bot: A typology to guide research and policy," *Policy Internet*, vol. 12, no. 2, pp. 225–248, Jun. 2020.

[84] S. Cresci, F. Lillo, D. Regoli, S. Tardelli, and M. Tesconi, "FAKE: Evidence of spam and bot activity in stock microblogs on Twitter," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 12, 2018, pp. 580–581.

[85] P. D. Turney, "Coherent keyphrase extraction via web mining," Assoc. Comput. Mach., San Francisco, CA, USA, Tech. Rep. NRC 46496, 2003.

[86] S. Siddiqi and A. Sharan, "Keyword and keyphrase extraction techniques: A literature review," *Int. J. Comput. Appl.*, vol. 109, no. 2, pp. 18–23, Jan. 2015.

[87] K. S. Hasan and V. Ng, "Automatic keyphrase extraction: A survey of the state of the art," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 1262–1273.

[88] Z. A. Merrouni, B. Frikh, and B. Ouhbi, "Automatic keyphrase extraction: A survey and trends," *J. Intell. Inf. Syst.*, vol. 54, pp. 391–424, Apr. 2019.

[89] A. S. Singh and C. S. Tucker, "Investigating the heterogeneity of product feature preferences mined using online product data streams," in *Proc. 41st Design Autom. Conf.*, Aug. 2015, pp. 1–26.

[90] M. S. Mubarok, Adiwijaya, and M. D. Aldhi, "Aspect-based sentiment analysis to review products using Naïve Bayes," in *Proc. AIP Conf. Proc.*, 2017, Art. no. 020060.

[91] X. Chen, Y. Xue, H. Zhao, X. Lu, X. Hu, and Z. Ma, "A novel feature extraction methodology for sentiment analysis of product reviews," *Neural Comput. Appl.*, vol. 31, no. 10, pp. 6625–6642, Oct. 2019.

[92] K. Pearson, "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philos. Mag.*, vol. 50, no. 320, pp. 157–175, 1900.

[93] M. Paquot and Y. Bestgen, "Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction," in *Corpora: Pragmatics discourse*. Leiden, The Netherlands: Brill Rodopi, 2009, pp. 247–269.

[94] M. A. Palomino and T. Wuytack, "Unsupervised extraction of keywords from news archives," in *Lang. Technol. Conf.* Berlin, Germany: Springer, 2009, pp. 544–555.

[95] P. Rayson, "Corpus analysis of key words," in *The Encyclopedia Application Linguistics*. Oxford, U.K.: Wiley, 2012.

[96] I. Ahmad, D. Tang, T. Wang, M. Wang, and B. Wagan, "Precipitation trends over time using mann-Kendall and Spearman's rho tests in swat river basin, Pakistan," *Adv. Meteorol.*, vol. 2015, pp. 1–15, Dec. 2015.

[97] D. Sharma, B. Kumar, and S. Chand, "A trend analysis of machine learning research with topic models and mann-Kendall test," *Int. J. Intell. Syst. Appl.*, vol. 11, no. 2, pp. 70–82, Feb. 2019.

[98] S. Malakar, S. Goswami, and A. Chakrabarti, "An online trend detection strategy for Twitter using mann–Kendall non-parametric test," in *Proc. Ind. Interact. Innov. Sci., Eng. Technol.* Singapore: Springer, 2018, pp. 185–193.

[99] J. K. Brooks, N. Bashirelahi, and M. A. Reynolds, "Charcoal and charcoal-based dentifrices: A literature review," *J. Amer. Dental Assoc.*, vol. 148, no. 9, pp. 661–670, Oct. 2017.

[100] R. Hosadurga, V. A. Boloor, S. N. Rao, and N. MeghRani, "Effectiveness of two different herbal toothpaste formulations in the reduction of plaque and gingival inflammation in patients with established gingivitis—A randomized controlled trial," *J. Traditional Complementary Med.*, vol. 8, no. 1, pp. 113–119, Jan. 2018.

[101] A. Timoshenko and J. R. Hauser, "Identifying customer needs from user-generated content," *Marketing Sci.*, vol. 38, no. 1, pp. 1–20, Jan. 2019.

[102] C.-S. Wu, C.-J. Kuo, C.-H. Su, S. Wang, and H.-J. Dai, "Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records," *J. Affect. Disorders*, vol. 260, pp. 617–623, Jan. 2020.

[103] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, pp. 37–46, Apr. 1960.

[104] G. Hripcsak and A. S. Rothschild, "Agreement, the F-measure, and reliability in information retrieval," *J. Amer. Med. Informat. Assoc.*, vol. 12, no. 3, pp. 296–298, 2005.

[105] C. Grouin, S. Rosset, P. Zweigenbaum, K. Fort, O. Galibert, and L. Quintard, "Proposal for an extension of traditional named entitites: From guidelines to evaluation, an overview," in *Proc. 5th Linguistics Annotation Workshop*, 2011, pp. 92–100.

[106] L. Deleger, Q. Li, T. Lingren, and M. Kaiser, "Building gold standard corpora for medical natural language processing tasks," in *Proc. AMIA Annu. Symp. Proc.*, 2012, p. 144.

[107] A. Brandsen and S. Verberne, "Creating a dataset for named entity recognition in the archaeology domain," in *Proc. LREC*, 2020, pp. 4573–4577.

[108] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 1, pp. 159–174, Mar. 1977.

[109] S. Hurtado, P. Ray, and R. Marculescu, "Bot detection in reddit political discussion," in *Proc. 4th Int. Workshop Social Sens.*, 2019, pp. 30–35.

[110] S. Raju, P. Pingali, and V. Varma, "An unsupervised approach to product attribute extraction," in *Proc. Eur. Conf. Inf. Retr.* Berlin, Germany: Springer, 2009, pp. 796–800.

[111] S. Raju, P. Shishtla, and V. Varma, "A graph clustering approach to product attribute extraction," in *Proc. IICAI*, 2009, pp. 1438–1447.

[112] N. Jakob and I. Gurevych, "LRTWIKI: Enriching the likelihood ratio test with encyclopedic information for the extraction of relevant terms," in *Proc. Workshop, User Contributed Knowl. Artif. Intell.*, 2009, pp. 3–8.

[113] L. M. Hamilton and J. Lahne, "Fast and automated sensory analysis: Using natural language processing for descriptive lexicon development," *Food Qual. Preference*, vol. 83, Jul. 2020, Art. no. 103926.

[114] N. Pandis, "The chi-square test," *Amer. J. Orthodontics Dentofacial Orthopedics*, vol. 150, no. 5, pp. 898–899, 2016.

[115] I. S. Bedmar, P. Martínez, and M. H. Zazo, "Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)," Assoc. Comput. Linguistics, Stroudsburg, PA, USA, Tech. Rep. S13-2056, 2013.

[116] L. Aiello, G. Petkos, C. Martin, and D. Corney, "Sensing trending topics in Twitter," *IEEE Trans. Multimeida*, vol. 15, no. 6, pp. 1268–1282, Oct. 2013.

[117] D. Gayo-Avello, "No, you cannot predict elections with Twitter," *IEEE Internet Comput.*, vol. 16, no. 6, pp. 91–94, Nov. 2012.

DAVID KILROY received the B.Sc. degree (Hons.) in computer science from Technology University Dublin, in 2019. He is currently pursuing the Ph.D. degree with University College Dublin. His research interests include machine learning, text mining, social media, and product development.

GRAHAM HEALY received the B.Sc. degree (Hons.) in computer applications, in 2008, and the Ph.D. degree in brain–computer interfaces, in 2012. He worked as a Postdoctoral Researcher with The University of British Columbia, from 2012 to 2013, and The Insight Centre for Data Analytics, Dublin City University, in 2013, where he later became a Research Fellow, in 2017, and then became an Assistant Professor in computing, in 2019. He is currently an Assistant Professor with the School of Computing, Dublin City University. His research interests include the ways computerized systems can automatically detect things from people using signals, i.e., bioelectric, social, and collaborative, then do something useful with that information, and mix of basic-research with a practical focus on developing real-world applications.

SIMON CATON received the B.Sc. degree (Hons.) in computer science, in 2005, and the Ph.D. degree in computer science, in 2010. He worked as a Postdoctoral Researcher with the Karlsruhe Institute of Technology, Germany, from 2010 to 2014, and a Lecturer in data analytics with the National College of Ireland, between 2014 and 2019. He is currently an Assistant Professor with the School of Computer Science with University College Dublin. His research interests include the applications of machine learning across the domains of social media, quantum computing, and parallel and distributed computing.