Detecting Pain Points from User-Generated Social Media Posts Using Machine Learning

Detecting Pain Points from User-Generated Social Media Posts Using Machine Learning

ABSTRACT

Artificial intelligence, particularly machine learning, carries high potentials to automatically detect customers' pain points, which is a particular concern expressed by the customer that the company can address. However, unstructured data scattered across social media makes detection a non-trivial task. Hence, to help firms gain deeper insights into customers' pain points, we experiment with and evaluate the performance of various machine learning models to automatically detect pain points and pain point types for enhanced customer insights. Our data consist of 4.2M user-generated tweets targeting 20 global brands from five different industries. Among the models we train, neural networks show the best performance at overall pain point detection with an accuracy of 85% (F1 score = 0.80). The best model for detecting five specific pain points was RoBERTa 100 samples using SYNONYM augmentation. With this study, we add another foundational building block of machine learning research in marketing academia through the application and comparative evaluation of machine learning models for natural language-based content identification and classification. In addition, we suggest that firms use *pain point profiling*, a technique for applying subclasses to the identified pain point messages to gain a deeper understanding of their customers' concerns.

Keywords: Marketing; artificial intelligence; AI; machine learning; customer insight; user-generated contents; UGC; pain points

1 INTRODUCTION

Obtaining customer insights—understanding customer needs, wants, and problems (Price et al., 2015), as well as relevant behaviors (Straker et al., 2021)—is considered a strategic asset for firms and underpin most marketing activities (Berger et al., 2020; Price & Wrigley, 2016). Obtaining deeper customer insights inform "the creation of offerings that are most aligned with their customer's needs and preferences." (Gupta et al., 2020, p. 27). Without a detailed understanding of customers' underlying needs and problems, it is hard to undertake effective marketing actions (Y. Liu et al., 2020). However, traditionally, it has been difficult for most firms to generate customer insights, mainly due to a lack of sufficient data and the firms' lack of ability to analyze them effectively (Price & Wrigley, 2016; Said et al., 2015).

There are new opportunities in this regard. Customers have increased their presence in the online landscape and leave a large degree of footprints in various forms as trace data (Freelon, 2014), including comments, expressions of interest, product reviews, or sharing of ideas. It is now possible to leverage these data by applying cutting-edge technologies (Reisenbichler et al., 2021) and generating deep customer insights (Antons & Breidbach, 2018; Berger et al., 2020). Towards these opportunities, this study offers understandings of how firms can collect and analyze users' Twitter posts (tweets) to automatically identify and categorize various types of customers' pain points, a particular type of customer insight (Homburg & Fürst, 2007; Rawson et al., 2013; B. Wang et al., 2016) that exemplifies how digital analytics can be leveraged to better understand the customers (Gupta et al., 2020). Artificial intelligence (AI), machine learning (ML) in particular, is an essential tool to this end.

We define a pain point as *an identifiable problem that the customers of a company have experienced that can be addressed by the company* (Homburg & Fürst, 2007; Rawson et al., 2013; B. Wang et al., 2016). For example, a customer's expression of "Company XYZ sucks!", even though critical of the company, does *not* express a pain point as it offers no actionable insight. However, a complaint like “I had to wait for 20 minutes over the phone to reach the customer service representative” does, because it reveals a specific problem of a long waiting time that the company can address (Handfield & Steininger, 2005; Rawson et al., 2013; B. Wang et al., 2016).

The automatic detection of customer-generated messages that contain a pain point is called *pain point detection*. Pain point detection is particularly relevant for managing customer relationships in the social media era (Malthouse et al., 2013), where customers’ pains are not physical but experiential and emotional (B. Wang et al., 2016). Customers’ disappointment or psychological gaps induce the pains because their expectations are not met through the offerings or during the customer journey (Kranzbühler et al., 2019; Rawson et al., 2013). Understanding pain points offers firm insights on customers’ critical issues, primary interests, and emerging demands for various offerings (Homburg & Fürst, 2007; Kranzbühler et al., 2019; B. Wang et al., 2016), thereby being instrumental for achieving a higher degree of market orientation (Kohli & Jaworski, 1990).

The fast development of information technology (IT) and communication technologies (ICT) has enabled customers to post a large volume of their concerns and expectation online (X. Liu et al., 2017). This user-generated content (UGC) is widely accepted to be a unique opportunity and valuable resource to generate customer insights (Cheng et al., 2021), including through identification and analysis of their pain points (Berger et al., 2020; X. Liu et al., 2017). However, the information abundance also makes pain point detection highly challenging (Balducci & Marinova, 2018; Kumar, 2018). Similar to most other problems that involve analyzing enormous datasets (Yang et al., 2021), understanding pain points by manually scrutinizing (big) data is an almost impossible task for humans (D. Cui & Curry, 2005). For example, how can one read millions of tweets, one by one, and determine whether they indicate pain points and what kind of customer insight they offer (Abu-Salih et al., 2018; Balducci &

Marinova, 2018)? The boundaries imposed by cognitive limitations in human capacity and the lack of tools to efficiently process pain points at scale make UGC, despite its high potential for customer insights, often left unused (Balducci & Marinova, 2018; Berger et al., 2020; D. Cui & Curry, 2005).

The application of and ML and, more specifically, its subfield of natural language processing (NLP) can help overcome these challenges and allow firms to gain customer insights at scale (Klapdor et al., 2014; Ma & Sun, 2020; Salminen et al., 2019). Large-scale, unstructured UGC be processed using ML techniques, which have versatile model structures and robust predictive performance (Amado et al., 2018; Ma & Sun, 2020; Rambocas & Pacheco, 2018). Even so, the application of these algorithmic developments in marketing is thus far being mainly focused on sentiment analysis (H. Li et al., 2022; Rambocas & Pacheco, 2018) and topic modeling (Amado et al., 2018; Reisenbichler & Reutterer, 2019) but not on pain point detection.

Thus, marketing research and application have great opportunities for deploying and applying innovative and advanced state-of-the-art methods that may generate superior insights that are well-applied in many other subject domains (Hartmann et al., 2019; Salminen et al., 2019). From an academic perspective, marketing as a research domain has been largely dependent on legacy methods that are simply incapable of dealing with the volume and complexity of data towards generating meaningful insights (X. Liu et al., 2017; Mustak et al., 2021). From the practical standpoint, even though there are very potent ML models and algorithms publicly available readily available to be trained for marketing problems, their vast potential to generate deeper customer insights from UGC remains largely untapped (Klapdor et al., 2014; Ma & Sun, 2020; Reisenbichler & Reutterer, 2019; Salminen et al., 2019). Specifically, there is a lack of research towards detecting pain points and further categorizing them into various groups of specific problems that firms can address (Mustak et al., 2021; Salminen et al., 2019).

Against this backdrop, *our study aims to develop the application of ML and NLP for generating customer insights through automated pain point analysis.* More specifically, we address the following research questions (RQs):

- **RQ1:** How can ML and NLP algorithms be successfully trained to generate customer insights through automatically detecting pain points from UGC?
- **RQ2:** Which machine learning model (type of algorithm) offers the best performance for pain point detection?
- **RQ3:** How can the best-performing model be trained further to identify the specific types of pains that the customers are facing?

Our goal is not to make technical contributions to the ML or NLP disciplines. Instead, we aim to develop, demonstrate, propose tools and offer recommendations to both marketing academics and practitioners on the applicability of the latest technological developments for generating customer insights from UGC via pain point analysis.

As tweets are one of the main forms of communication between brands and their customers, we collect and analyze 4.2 million customer-generated tweets targeted at 20 globally established brands from five different industries (Abu-Salih et al., 2018; X. Liu et al., 2017). This study makes three major contributions to the marketing literature deemed necessary by scholars (Berger et al., 2020; Ma & Sun, 2020; Mustak et al., 2021). First, *we demonstrate how ML can be used to generate customer insights*. Second, we compare various ML algorithms for automatic pain point detection, *showing that a neural network with transformer features yields the best performance*. This performance comparison serves as a valuable baseline for future research (Salminen et al., 2019). Third, our empirical analysis indicates that *customers mainly experience five types of pain points—product features or quality, service* *quality or failure, operational issues, customer services, and company's image*—thus offering further insights to marketing academics and practitioners (B. Wang et al., 2016). Going forward, firms can train the models with company-specific data, as necessary, to enhance the insights into specific pain points for specific customer populations.

The remainder of this article is organized as follows. First, we review and present related literature on customer insights, methods for pain point detection, and challenges of pain point detection. Then, we report our applied methodology in detail, including introducing ML classification models to the reader. Following that, we report the findings of our study, including algorithm selection, training, and performance evaluation, along with demonstrating their application. Finally, we offer deeper scrutiny into the performance of the best model, draw implications, identify limitations, and offer suggestions for future research.

2 RELATED WORK

2.1 Customer Insights

Customer insights are defined as an understanding of the various problems that customers face, including information about the customer's expressed and latent current and future needs (Price & Wrigley, 2016). By understanding their customers' experiences, desires, and expectations, businesses can benefit in a variety of ways, such as developing a more effective product strategy, determining how the company's actions affect its customers, executing more efficient marketing, or developing guidance toward greater differentiation from competitors (Berger et al., 2020; Macdonald et al., 2012; Price & Wrigley, 2016). However, understanding customers' experiences and how and why customers think and behave in certain ways remains a constant challenge for marketers (Price & Wrigley, 2016).

Customer insights can be classified into three broad categories according to their intended use: *instrumental*, *conceptual*, and *symbolic* (Macdonald et al., 2012; Said et al., 2015). Instrumental application entails the application of insight in specific, direct ways to resolve a specific dilemma that pertains to a current opportunity (Macdonald et al., 2012). Conceptual use entails utilizing customer insight for the purpose of enlightenment, influencing choices and behaviors more indirectly than instrumental use without taking relatively immediate tangible action (Said et al., 2015). Symbolic use includes applying customer insight to justify and sustain previously held positions and using insight to justify subsequent actions. (Macdonald et al., 2012; Said et al., 2015). Identifying and addressing customers' pain points directly connect to both instrumental and conceptual insights.

Traditional techniques for inferring customer insights tend to rely on survey methods and interviews (Griffin & Hauser, 1993; Johnson, 2007). The following section discusses these methods, their shortcomings, and the potential for automation in this problem setting. Essentially, people tend to voluntarily post their thoughts and experiences about brands (Rooderkerk & Pauwels, 2016; M. Zhang et al., 2011)—also known as "expressive individuality" (Weinberg et al., 2013)—that offers considerable opportunities for inferring customer insights and for understanding the relationship between customers and brands.

2.2 Methods for Identifying Customers' Pain points

Traditionally, to understand customers' pain points, and thus to generate customer insights, academics and managers have relied on various manual methods, as we summarize below (Macdonald et al., 2012):

- **Interviews:** Interviewing the customers to understand their pain points and thus to generate insights generally involve using open-ended questions to capture answers that reflect users' sentiments about a product or service (Schaffhausen & Kowalewski, 2015). Sample sizes vary across studies. For instance, Griffin and Hauser (1993) suggest that having one-hour interviews with 20–30 participants can elicit 90–95% of user needs.

- **Observations:** This method involves active user participation or detached examination of user practices and is generally concerned with what people do rather than what they say about their needs. It does not require users' conscious awareness of their needs to capture them (Patnaik & Becker, 1999).

- **Focus groups:** This method involves bringing customers together to discuss pre-determined topics (Griffin & Hauser, 1993). Discussions are guided by facilitators who typically follow a script but also probe participants for more profound answers.

- **Cross-sectional Surveys:** Various probability sampling methods are used to select the data. Typically, researchers invite users to respond to open- or close-ended questions using Likert scales or other means of categorization. Data analysis is performed using statistical techniques, such as conjoint analysis (Chen et al. 2019).

However, the existing literature mentions various challenges associated with these manual methods. They include budgetary constraints, time restrictions, difficulties in characterizing needs, small samples, and human biases**.** For instance, manual gathering and processing of data are often expensive (Kühl et al., 2019; Schaffhausen & Kowalewski, 2015), requiring subject-matter experts at all stages of the research process instrument design to data analysis (Y. Wang et al., 2018). Moreover, the manual methods are labor-intensive, time-consuming, and may require weeks in the field observing behavior or interacting with participants. Besides, various manual methods to detect pain points face the risk of researcher bias—for example, over-identification with research participants or pre-existing beliefs — threatening objectivity and compromising results (Kühl et al., 2019; Schaffhausen & Kowalewski, 2015; Y. Wang et al., 2018).

Against these challenges, automated techniques provide a notable opportunity to understand customer needs (Salminen et al., 2019) and understand the relationship between firms and customers (Libai et al., 2020). The large volume of data with information on customers’ pain points is often readily available online, which could generate deep customer insights (Y. Liu et al., 2020). Moreover, once the ML-based methods are developed and properly implemented, generating customer insights becomes relatively fast and much less resource-intensive (Ma & Sun, 2020; Mustak et al., 2021).

As a notable example, Lee and Bradlow (2011) used text mining techniques to process the text from online camera reviews and identified product attributes and attribute dimensions among brands and market segments. By comparing user reviews and expert reviews, the researchers developed insights on how the two groups valued different camera features. Wang et al. (2018) applied deep learning to identify latent user needs for new product design. Zhou et al. (2020) combined ML techniques, including sentiment analysis, to examine online product reviews within a product ecosystem (Amazon), labeling their problem as *customer needs analysis*. Liu, Dyzabura, and Mizik (2020) propose a neural network model for predicting brand attributes in online images, reporting high agreement between human raters and the model, as well as with consumer brand perceptions and survey data. Wang et al. (2020) investigated issue-tracking systems in open-source software communities and proposed an automated system for consolidating community opinions on usability issues. Kühl et al. (2020) called the problem *needmining*, and classified customer needs in the automotive industry using three types of classifiers (none of which included transformers). Conceptually, these problem definitions are similar, and they tend to deal with extracting customer concerns from unstructured textual data in the wild (typically in social media or other online platforms). More broadly, the problem of pain point detection deals with *generating customer insights from the voice of customers*.

2.3 Challenges for Automatic Pain point Detection

Despite the advantages of automated methods, there are several challenges when applying ML for pain point detection. Based on the literature, these include at least the following:

- **Noisy or low-quality data**: Working with prodigious amounts of unstructured data involves a vast number of irrelevant or uninformative samples. For example, online product reviews may contain comments that do not address specific needs. This noise needs to be removed, and pertinent information must be extracted before training the algorithm (Timoshenko & Hauser, 2019; Y. Wang et al., 2018; Zhou et al., 2020). The adage "quality versus quantity" is extremely pertinent to text analytics. If the original data are poor in quality, then an ML model will not perform well (i.e., "garbage in, garbage out").
- **Semantic ambiguity:** Many ML algorithms, particularly supervised ML, require large amounts of data before generating usable results. In their study, Kuhl et al. (2019) sought to circumvent this process by using unsupervised ML to quantify user needs. However, their testing of multiple unsupervised clustering possibilities did not yield results that made semantic sense. Consequently, they concluded that their dataset was insufficient for an unsupervised approach and devised a process for supervised ML instead.
- **Lack of standards:** Humphreys and Wang (2018) argue that there is no standard set of methods, steps of inclusion and exclusion, sampling, and dictionary development and validation in automatic text analysis of UGC. They also suggest that the failure to integrate linguistic theory into automated text analysis limits the field and prevents knowledge about "the multiple dimensions of language that can be used to measure user, thought, interaction and culture" (p. 1275).
- **Social desirability bias:** Automated text analysis tends to rely on publicly available data in the form of tweets, Facebook postings, or online reviews. Because these data are public,

users may not feel comfortable sharing information that they feel is socially unacceptable or undesirable. In some cases, this information might be relevant to data analysis, and its absence may skew research findings (Humphreys & Wang, 2018).

- **Role of physicality:** There are situations when observing behavior is the only way to study a phenomenon. Automated methods, conducted remotely, cannot capture how users interact tactilely with products or navigate space in user environments (Humphreys & Wang, 2018). As such, textual data collected in the wild has intrinsic limitations for customer understanding.
- **Need for human involvement:** Although ML is generally considered to be significantly less time- and labor-intensive than manual methods, the amount of work these methods require is often underestimated. However, in reality, a significant degree of effort and time is often necessary to prepare and apply ML-based methods. Automated methods demand human-labeled inputs before algorithms in supervised ML can be trained (Zhou et al., 2020). Researcher involvement in design, modification, and interpretation may be necessary at other points, such as dictionary development and validation. In some cases, manual data assessment is necessary after the automatic methods are completed.

3 METHODOLOGY

Despite its potential, the application of ML is still, in many regards, in the early stages in marketing research (Ma & Sun, 2020; Salminen et al., 2019), although, as noted in the literature review, there is prior work in this area. Hence, we keep the technical details reporting simple to not deter the interest of the broader marketing readership. This also helps the interested reader gain a general understanding of ML that we believe will continue gaining prominence as a methodological choice in the marketing domain and enabling new research possibilities (Balducci & Marinova, 2018; Davis et al., 2013).

3.1 Use of Machine Learning to Generate Customer Insights

The widely used definition of ML is offered by Mitchell (1997): *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."* ML is a broad term that refers to a number of computer-based data-mining and interpretation techniques for uncovering complex patterns, especially in large and complex datasets (Mohri et al., 2018; Shalev-Shwartz & Ben-David, 2014), with a particular drive to deriving insights for classification, prediction, and decision-making purposes (G. Cui et al., 2006).

There are two main types of learning mechanisms for the machines within the ML field, i.e., computers and algorithms: unsupervised and supervised (Kotsiantis et al., 2007; Salian, 2018). In unsupervised ML, models are developed based on unlabeled data that the algorithm tries to make sense of by extracting features and patterns on its own. In contrast, in supervised learning, a "machine" (statistical model) is trained using "labeled" data (Asiri, 2018; Kotsiantis et al., 2007). The training datasets are already labeled with correct answers—they contain both input and output parameters. Accordingly, supervised learning is ideally suited to problems with a set of available reference points or ground truth to train the algorithm with (Ray, 2017). For example, a training dataset of animal images would mean each photo was pre-labeled as dog, turtle, or koala. The algorithm is then assessed based on how well it can distinguish new photos of koalas and turtles (Salian, 2018).

Classification is a major part of ML—it allows the algorithm to automatically determine which class (a.k.a. group) an observation belongs to (Kotsiantis et al., 2007; Yiu, 2019). Spam detection in email service providers is a typical classification problem; this is a binary classification since there are only two classes—spam and non-spam (Yiu, 2019). An example of supervised ML-based classification is an algorithm that predicts the price of an apartment in San Francisco based on square footage, location, and proximity to public transportation (Asiri, 2018; Salian, 2018). This ability to precisely identify observations is beneficial for a variety of business applications, such as predicting whether a specific customer will purchase a product or whether a given loan will default (Asiri, 2018; Yiu, 2019). The process of applying supervised ML to a real-world problem is illustrated in Figure 1 (Kotsiantis et al., 2007).

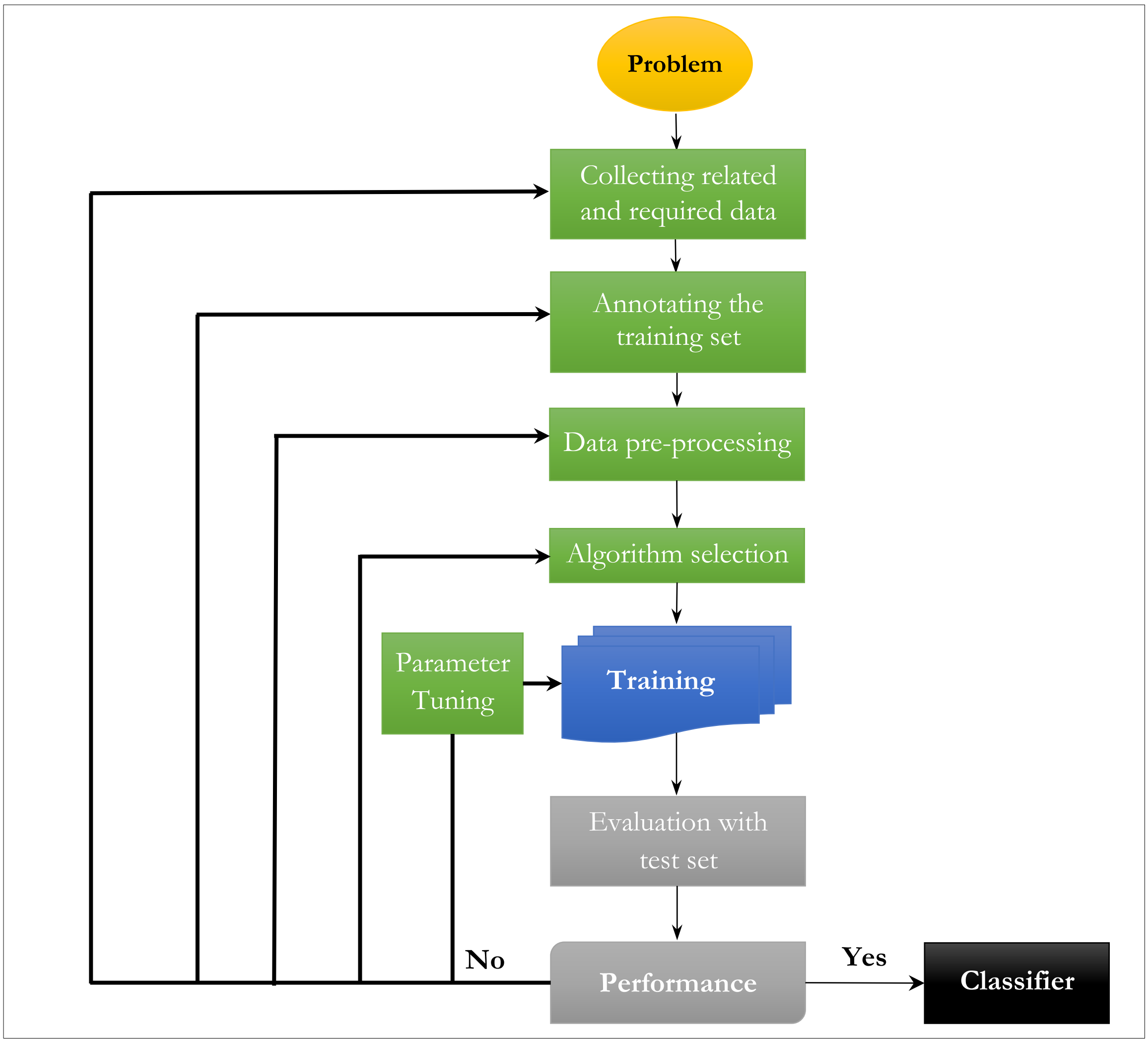


Figure 1: The process of supervised machine learning (based on Kotsiantis et al., 2007).

The first step is to collect the dataset, for example, customers' tweets or product reviews. Second, a sub-set (training set) is separated and annotated with the correct answers from the collected data. This is typically a human-performed operation, for instance, manually identifying the tweets that contain pain points and then classifying the pain points further. The preparation and preprocessing of data is the third stage. The next step is deciding which learning algorithm to use. Following that, the algorithm must be trained by inputting data, which creates a classifier (i.e., classification model). This often requires the setting of hyperparameters, noted as parameter tuning. Once the classifier's predictive accuracy is deemed satisfactory, the classifier (which maps unlabeled instances to classes) is ready for use on a regular basis (Kotsiantis et al., 2007), i.e., detecting pain points in our case.

3.2 Data Collection

We train our algorithm to identify a wide range of pain points for deep customer insights (Salian, 2018; Salminen et al., 2019). In doing so, our starting point was to select companies from different industries so that our algorithm develops the capability to identify a diverse set of customers' complaints and viewpoints. Following X. Liu et al. (2017), we chose five industries based on the Global Industry Classification Standard—consumer services, food and beverage, retailing, apparel and footwear, and electronics industries. We selected four representative brands that rank within the Forbes Global 2000 list for each industry. The industries and respective companies are presented in Table 1.

Table 1: The industries and representative companies included in this study. These represent well-known global brands

	Industry				
	Consumer Services	**Food and Beverage**	**Retailing**	**Apparel & Footwear**	**Electronics**
Companies	FedEx	Coca-Cola	Amazon	Adidas	Fitbit
	Marriot	McDonald's	Macy's	Gap	Nintendo
	Netflix	Nestle	Tesco	Nike	Samsung
	Uber	Starbucks	Walmart	Puma	Sony

For data collection, Twitter is a prominent source of datasets. While Twitter does not directly provide datasets of user messages (tweets) to the public, researchers can use Twitter's application programming interface (API) to retrieve tweets that match predefined conditions.

We mined tweets aimed directly at the brands to maximize the relevance of the collected tweets for the research goal. Using Twitter's API, we applied the pattern "@company" (replacing the word "company" with a brand name each time) to only gather tweets specifically addressed to the companies' Twitter accounts, which enabled us to retrieve approximately 200,000 tweets per brand, i.e., 4.2 million tweets in total (precisely 4,209,101), excluding any retweets.

3.3 Annotation of Training Data

We used stratified sampling to retrieve 2,000 tweets among the 4.2 million tweets collected to build a pain point detection dataset. The stratification factor was the brand name, so each of the 20 brands had 100 samples in the dataset. Thus, we ensured that our training set contained an equal representation of the brands. Based on our initial estimate, 2000 tweets were deemed appropriate to balance the manual labor required for the annotation and the provision of an adequate number of samples for training the algorithm.

Two researchers independently annotated the tweets, i.e., whether they contained any pain points or not. The tweets that contained a pain point were marked with "Y", and the others were marked with "N". In other words, the dataset was created for a binary classification task —to train classifiers for detecting if a tweet contains a pain point or not.

We iteratively annotated the dataset – after a sample of the tweets was annotated independently by two researchers, a third researcher checked for the disagreements—for this, we computed Cohen's kappa (*k*) (McHugh, 2012). We then manually examined and discussed each disagreed instance to come to a conclusion. After that, we moved forward with annotating the next sample of the dataset.

In the first round, where each annotator annotated 150 random samples, the agreement was only 78% (*k*=0.54, indicating moderate agreement). The disagreed instances were resolved one-by-one among the researchers, coming up with a consensus on the true label in each case. The second round consisted again of 150 randomly selected samples (excluding those previously tested). The second-round agreement was 85% (*k*=0.69), moving from moderate to substantial agreement. Again, we repeated the approach of investigating the disagreed instances and the reasoning to reconcile the views.

In the third iteration, conducted as before, we achieved an overall inter-rater agreement of 90.3% (*k*=0.81), indicating almost perfect agreement. Again, the found disagreements were solved by discussion to formulate labels that contain the least possible degree of subjectivity. The reader should note that the nature of the problem implies that it is unlikely that one would be able to create perfect agreement, i.e., there is always some degree of subjectivity whether a message contains a pain point or not—this is referred to as "inherent subjectivity" in annotation tasks (Alonso, 2015; Alonso et al., 2015; Salminen et al., 2018).

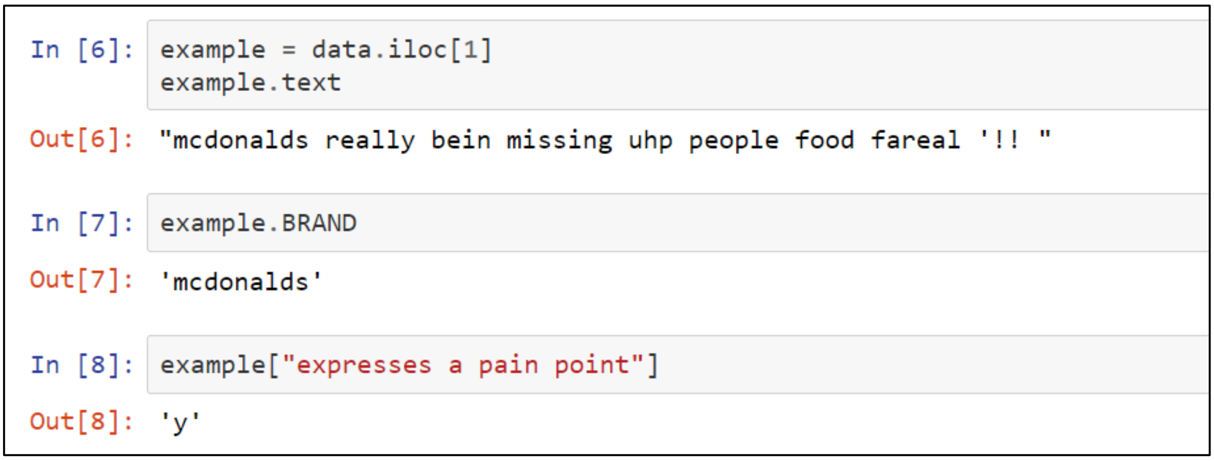
The discussion and examples that offer further insights on how the annotation evolved are available from the authors to any interested party. After these three rounds, one of the researchers continued to annotate further samples until a total of 2,000 annotations were reached. This process resulted in a dataset where 656 (32.8%) tweets had a paint-point (i.e., positive class), and 1,344 (67.2%) tweets did not have a pain point (i.e., negative class). Separating the positive samples from the negative ones will constitute the binary classification task (i.e., pain point detection).

3.4 Data Preprocessing and Exploration

Data preprocessing, which includes cleaning and exploring the data, is essential in ML projects, particularly in this type of case where the primary data consists of colloquial language. In addition, together with the contents, tweets contain other extraneous information, such as the number of likes and retweets. Therefore, we conducted data cleaning to eliminate irrelevant text content, including extra white space characters, non-alphabetic characters, and stopwords (i.e., words that have no actual meaning in the text, like "and," "the," and "or"). This ensured that our data set was implementable in all the range of machine learning models that we employed.

Next, this project's data preparation process is vital since computers ultimately can process only numerical data. This means we had to convert the tweets, which are composed mainly of text, into numbers. We used the *Term Frequency-Inverse Document Frequency* (TF-IDF) algorithm to convert the tweet into the numerical format for easier consumption by the learning algorithms. TF-IDF assigns scores to each word based on how common they are in a specific article and how uncommon they are across all articles (Salminen et al., 2019). We also used the Bidirectional Encoder Representations from Transformers (BERT) transformer to get context-embeddings (Horev, 2018).

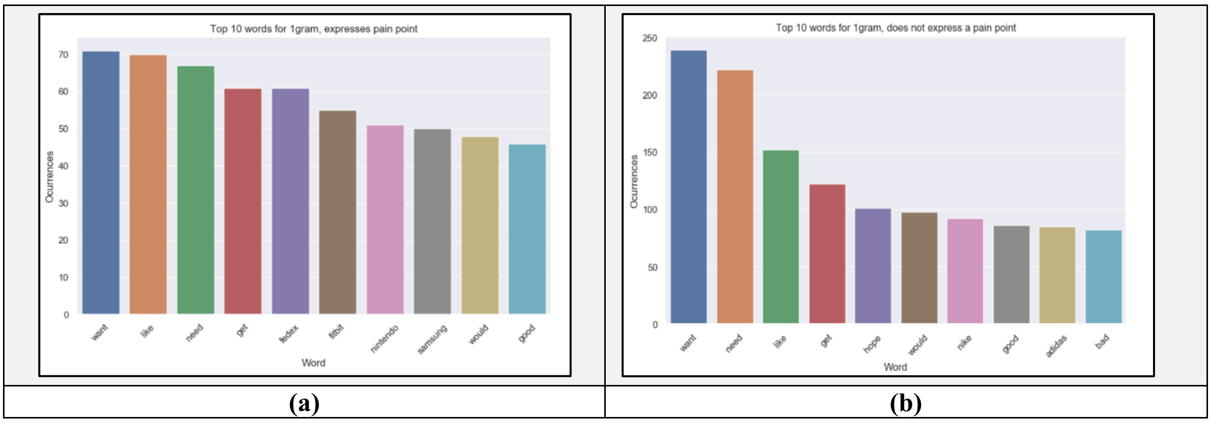
To explore the data, we picked up some tweets randomly to examine their characteristics in relation to the study. Figure 2 offers an illustration of a random exploration.



```
In [6]:  example = data.iloc[1]
         example.text

Out[6]:  "mcdonalds really bein missing uhp people food fareal '!! "

In [7]:  example.BRAND

Out[7]:  'mcdonalds'

In [8]:  example["expresses a pain point"]

Out[8]:  'y'
```

Figure 2: An illustration of data exploration using algorithms.

From Figure 2, the correlation between the target and the text is visible. The brand is also evident in the tweet, as well as the discomfort of the user and the specific pain point. The high colloquial nature of the customer’s language is also noticeable. This is a clear indication that the ML algorithms must deal with a wide range of informal and idiomatic texts to successfully capture pain points. Next, to further explore the data and analyze its aptness for the study, we developed visualizations of the top words and combinations of words concerning whether the tweets express a pain point or not. To do so, the first step was to use N-grams, which is a contiguous sequence of n items from a given sample of text or speech. Figure 3 shows the 10 top unigrams (N = 1) for tweets that express pain points or not.



(a) **(b)**

Figure 3: Unigrams from tweets that **(a)** express and **(b)** do not express pain points.

As can be seen from Figures 3a and 3b, some words are very common in both types of tweets. The words “want,” “like,” and “need” form the top three words in our dataset in tweets that have no pain points and in those that do, making the task more difficult. However, increasing to N = 2 (a string of two consecutive words) improves the signal somewhat, as shown in Figure 4. Nonetheless, the features (i.e., numerical representations) we infer from the data are more complex than uni- or bigrams, containing more information and sentences and semantic meaning (see Section 4.3 for explanation).

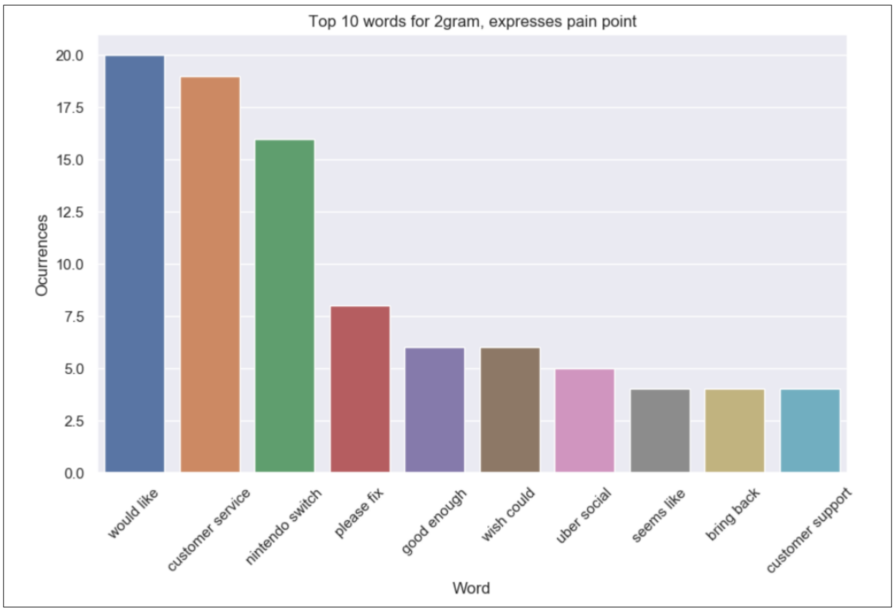


Figure 4: Bigrams (N = 2) of words from tweets that express pain points

Overall, the exploration of our data indicates that it is suitable for training the algorithms to detect customers' pain points to generate customer insights. However, the language's high degree of colloquiality and idiomatisms would require relatively advanced learning. For this, algorithm selection and hyperparameter setting are crucial steps.

We also identified an interesting statistical association in the dataset. The character length of tweets that mentioned a pain point (M = 144, SD = 68.89) was significantly higher compared to the tweets that did not express a pain point (M = 112, SD = 64.03), t(1999) = 9.923, $p < 0.001$. This implies that customers tend to be more vocal when expressing pain points than when engaged in other types of communication (see Figure 5). We also tested the number of retweets and replies, but these factors did not significantly differ between pain point and non-pain point tweets.

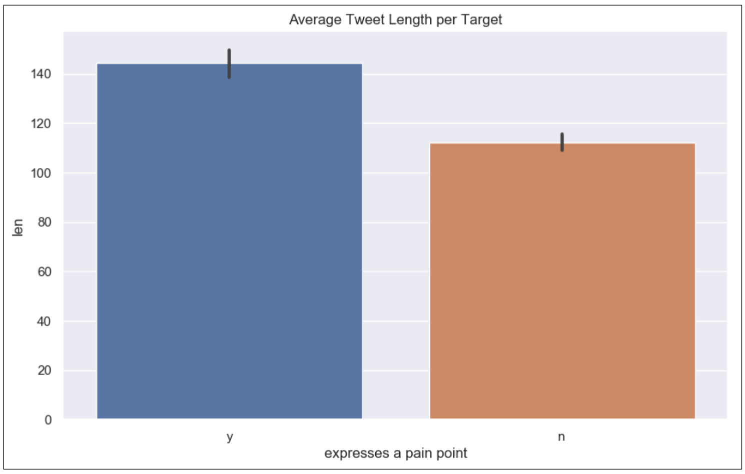


Figure 5: The length of tweets that express and do not express a pain point.

3.5 Algorithm Selection and Training

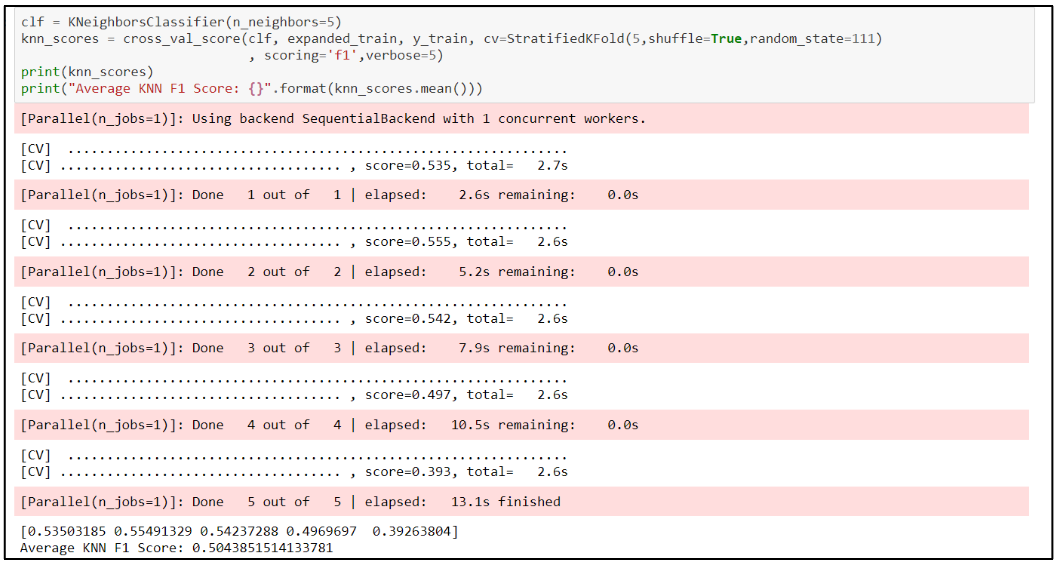
Data science and ML offer a plethora of classification algorithms (Ray, 2017; Yiu, 2019). This study selects six different models that are considered powerful for classification tasks (Asiri, 2018; Breiman, 2001; Kramer, 2013; Steinwart & Christmann, 2008). The models, along with their definitions, are presented in Table 2.

Table 2: Classification algorithms used in this study.

Classification Algorithm	Definition
K-Nearest Neighbors	Algorithms that presume similarities between the new case/data and existing cases and place the new case in the most similar category to the existing categories (Asiri, 2018; Kramer, 2013)
Random Forests	Algorithms that average the outcomes of a number of decision trees applied to various subsets of a dataset to improve the dataset's predictive accuracy (Breiman, 2001; Donges, 2019; Pawar, 2020)
XGBoost	Algorithms that implement gradient-boosted decision trees to provide classification (Brownlee, 2016; T. Chen & Guestrin, 2016; Reinstein, 2017)
Naïve Bayes	Algorithms based on the Bayes theorem that provide probabilistic classification, i.e., predict an object's probability (Rish, 2001)

Classification Algorithm	Definition
Support Vector Machines	Algorithms that plot each data object as a point in n-dimensional space, with the value of each function being the value of a specific coordinate, and perform classification by determining the hyperplane that distinguishes the groups (Steinwart & Christmann, 2008)
Neural Networks	Algorithms that are modeled loosely after the human brain and are used to model complex patterns in datasets using multiple hidden layers and nonlinear activation functions (Knocklein, 2019; Rojas, 2013)

Overall, with their rather high level of abilities to deal with natural human language, these six ML algorithms are expected to reflect the state-of-the-art performance for automatic pain point detection (Asiri, 2018; Breiman, 2001; Kramer, 2013; Steinwart & Christmann, 2008). To train the models, we pass the transformed tweets that have been converted into a table of numbers, along with a variable that indicates if the tweet expresses a pain point or not. As an illustration, Figure 6 illustrates KNN (k-nearest neighbors' algorithm) modeling.



```
clf = KNeighborsClassifier(n_neighbors=5)
knn_scores = cross_val_score(clf, expanded_train, y_train, cv=StratifiedKFold(5,shuffle=True,random_state=111)
                             , scoring='f1',verbose=5)
print(knn_scores)
print("Average KNN F1 Score: {}".format(knn_scores.mean()))
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.

[CV]  ................................................................
[CV] .................................... , score=0.535, total=   2.7s

[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    2.6s remaining:    0.0s

[CV]  ................................................................
[CV] .................................... , score=0.555, total=   2.6s

[Parallel(n_jobs=1)]: Done   2 out of   2 | elapsed:    5.2s remaining:    0.0s

[CV]  ................................................................
[CV] .................................... , score=0.542, total=   2.6s

[Parallel(n_jobs=1)]: Done   3 out of   3 | elapsed:    7.9s remaining:    0.0s

[CV]  ................................................................
[CV] .................................... , score=0.497, total=   2.6s

[Parallel(n_jobs=1)]: Done   4 out of   4 | elapsed:   10.5s remaining:    0.0s

[CV]  ................................................................
[CV] .................................... , score=0.393, total=   2.6s

[Parallel(n_jobs=1)]: Done   5 out of   5 | elapsed:   13.1s finished

[0.53503185 0.55491329 0.54237288 0.4969697  0.39263804]
Average KNN F1 Score: 0.5043851514133781
```

Figure 6: An illustration of KNN modeling. Based on the data and the parameters, each model tries to learn a representation that accurately separates tweets that contain pain points and tweets that do not

4 RESULTS

4.1 Preliminary Experiments

Firms need to deploy high-performing ML models to generate customer insights by automatically identifying their pain points from a large dataset. Our analysis shows that the Neural Network-based model obtained the highest F1 score in automatically detecting customers' pain points (see Table 3). In this comparison, we also included three heuristic baseline models: (a) *keyword-based classifier (KBC)*, (b) *sentiment-based classifier (SBC)*, and (c) *length-based classifier (LBC)*.

KBC works by detecting if a tweet contains one or several keywords (if it does, it is marked as pain point; if not, it is marked as non-pain point). **SBC** checks both the presence of keywords and the sentiment – if the tweet has one or more pain point keywords, AND its sentiment is negative, it is a pain point; otherwise, no (the sentiment was classified using the VADER sentiment analysis tool (Hutto & Gilbert, 2014)). **LBC** marks a tweet as pain point if the tweet has one or more pain point keywords, its sentiment is negative, AND its character count is 144 or more (this was the average length of pain point tweets in the dataset). The keywords were compiled by expanding the list provided by Kühl et al. (2020) on their article's GitHub page[1]; two of the researchers created a list of terms that seem to indicate a pain point with a presumably high probability.

As can be seen from Table 3, the performance of these models was somewhere in between the worst and the best classification algorithms. As these baselines are based on the notion of heuristics, it appears that some heuristics provide a degree of signal for pain point detection: most notably, the best baseline is SBC (F1 = 0.51). The general difficulty of keyword-based approaches include the following: (a) individual words and phrases, when used in context, might not indicate actual pain points and thus have a high chance of false negatives

[1] https://github.com/cran/needmining/blob/master/R/filterTweetNeedwords.R

and positives (Salminen et al., 2020), (b) they ignore the use of negatives (i.e., "i did not want anything else" includes "want" and the classifier would consider this a pain point even though the user indicates they are satisfied), (c) they are challenging to create comprehensively, i.e., to reflect all aspects of the analyzed phenomenon. Nonetheless, the list of keywords we used is provided in supplementary material on GitHub for further development and scrutiny of researchers: **[link hidden for anonymous review]**.

Table 3: F1 scores of the tested models. 'Heuristic' indicates that the model is based on hand-crafted rules (explained in the body text). 'Learning' indicates a statistical machine-learning model. The highest performance (Neural Network) is bolded.

Type	Classification Algorithm	F1 Score
HEURISTIC	Keyword-based classifier	0.31
LEARNING	Support Vector Machines	0.41
LEARNING	Random Forests	0.44
HEURISTIC	Keyword+sentiment+length	0.47
LEARNING	K-Nearest Neighbors	0.50
HEURISTIC	Keyword+sentiment	0.51
LEARNING	XGBoost	0.55
LEARNING	Naïve Bayes	0.59
LEARNING	**Neural Network**	**0.60**

For the model comparison, we applied the models' default Scikit-learn parameters. No hyperparameter tuning was made at this stage to enable a standard experimental comparison across the algorithms. For model validation, we split the data into training and testing. The models were validated using 5-fold stratified cross-validation (Krogh & Vedelsby, 1995), using the original training set (this results in 5 validation sets and 5 training sets).

After training all six models with the same annotated dataset, we assessed their performance through their F1 scores. The F1 score can be interpreted as a weighted average of the "precision" and "recall." *Precision* is the ratio of correctly predicted positive observations to the total predicted positive observations. *Recall* is the ratio of correctly expected positive observations to all observations in the actual class. The F1 score is calculated as below:

*2 * (precision * recall) / (precision + recall)*

Thus, in the F1 score, the relative contribution of precision and recall to the F1 score are equal. The advantage of the F1 score is it incorporates both *precision* and *recall* into a single metric, and therefore gives a rather fair assessment of performance in the case of class imbalance (scikit-learn, 2020).

The best model was NN (F1 = 0.60). We used TensorFlow (https://www.tensorflow.org/), an open-source ML library, to create the NN architecture. The NN had one fully connected (dense) layer. We used the rectified linear activation function (ReLU) that selectively activates neurons in order to maintain a high computational efficiency (Y. Li & Yuan, 2017). We also used the dropout technique (parameter value = 0.5) to prevent overfitting (Srivastava, 2013). The optimizer was Adam (Kingma & Ba, 2015), an extension of stochastic gradient descent commonly used to optimize NN models. The network was trained over 5 epochs using the batch size of 8.

To further improve the performance of the NN model, we applied hyperparameter optimization (Agrawal, 2020) and tested it against the test set of our data, which comprised 20% of the dataset that had not been used during the training and was therefore unfamiliar to the model. The hyperparameter optimization was done using grid search (Bergstra & Bengio, 2012) and 5-fold stratified cross-validation. Since the selected model was NN, the parameters optimized included (a) the number of neurons, (b) the number of layers, (c) batch size, and (d) the number of epochs, optimized using the *GridSearchCV* library (Pedregosa et al., 2011). The final hyperparameters were then tested against the test set. In other words, we used and training set, a hold-out test set, and used cross-validation to generate several validation sets. These validation sets were shuffled but with a fixed random seed (see (Sugimura & Hartl, 2018)).

After the optimization, the macro-average F1 score of the model was **0.66**, which is an improvement of 9.3% from the non-optimized NN. The overall accuracy was **76.5%,** translating to roughly *three out of four* tweets being correctly classified.

To check the specificities of the model's strength in detecting customers' pain points from their natural language-based expressions, we tested its performance against each of the 20 brands in our sample. Tweets targeting Fitbit, Nintendo, and FedEx offered the best precision in terms of ML to detect pain points automatically. In contrast, the model was not capable at all of determining whether tweets targeted toward Puma or Gap contained pain points (see Figure 7).

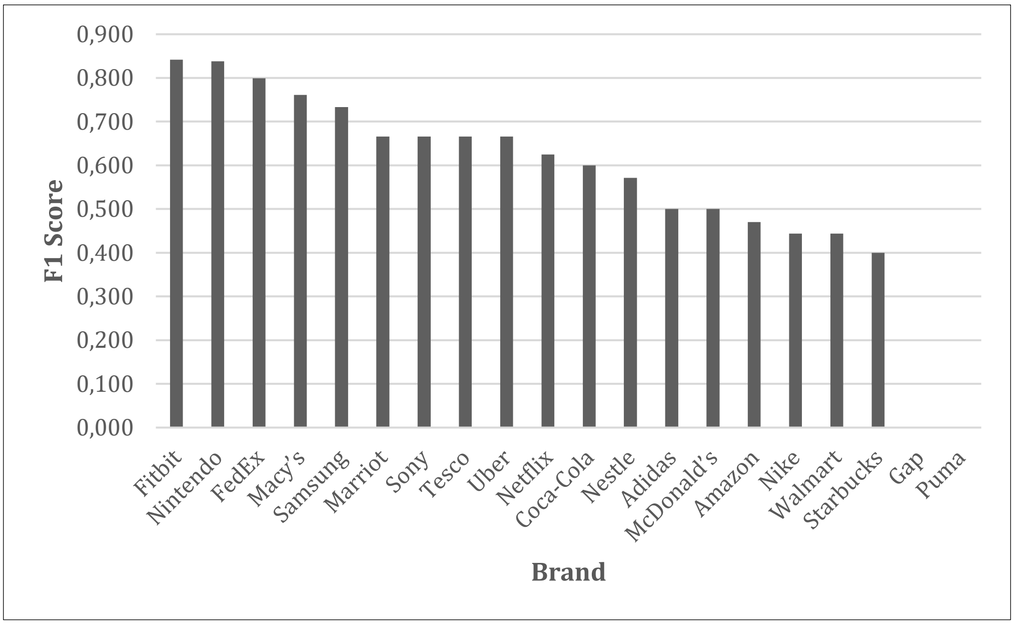


Figure 7: Brand-specific F1 score of neural network-based model for detecting customers' pain points.

We further investigated why the model performs well for particular brands and worse for others. Our investigation indicates that for low-performing brands, the actual body texts of the tweets often do not the name of the company or brand, thus lowering the model's predictive performance.

4.2 Interpreting Model’s Decisions

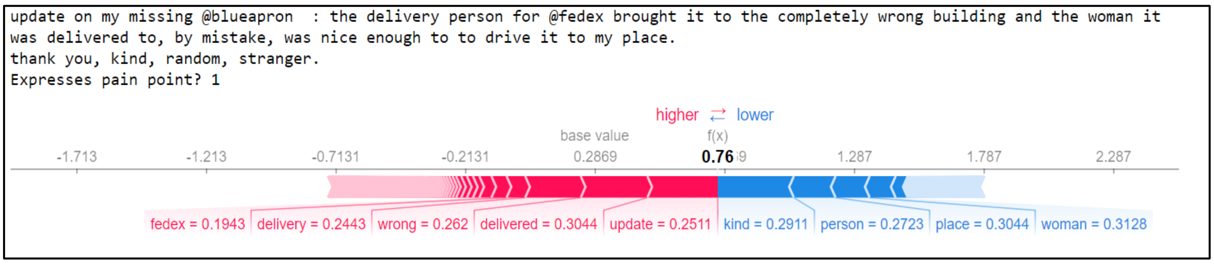
Next, we show the application of the model for pain point detection. In a ML model, the TF-IDF features are interpretable by humans (Salminen et al., 2019). Hence, based on TF-IDF features, we show different tweets, offer our interpretation, and then present the result found through the newly trained NN model. If the algorithm can detect a pain point, then it shows the value “1.” If it interprets that the tweet does not contain a pain point, then the value is “0”. We show five examples for the reader to understand how the results can be interpreted.

Example 1:

Original tweet: “update on my missing @blueapron : the delivery person for @fedex brought it to the completely wrong building and the woman it was delivered to, by mistake, was nice enough to drive it to my place. thank you, kind, random, stranger.”

Our interpretation: The customer is complaining about her/his package being delivered to the wrong address by the delivery services company. This tweet expresses a pain point as it clearly shows the problem and offers actionable insight—improving delivery performance.

Assessment performed by neural network: 1 (Contains pain point)



```
update on my missing @blueapron  : the delivery person for @fedex brought it to the completely wrong building and the woman it
was delivered to, by mistake, was nice enough to to drive it to my place.
thank you, kind, random, stranger.
Expresses pain point? 1
```

Figure 8: Automated detection of customer’s pain points. Assessment—positive. The NN model has been able to make the correct assessment.

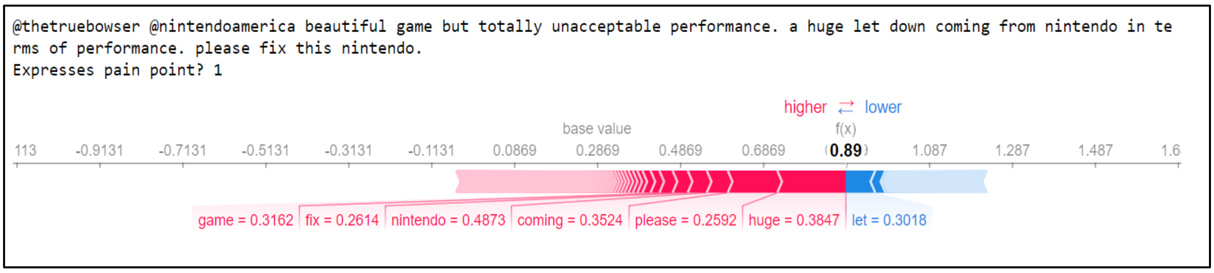
Final outcome: The NN model has been able to make the correct assessment.

Example 2:

Original tweet: "@thetruebowser @nintendoamerica beautiful game but totally unacceptable performance. a huge let down coming from Samsung27 in terms of performance. please fix this Samsung27."

Our interpretation: The customer is unhappy about the performance of Bowser, a fictional character and the main antagonist of Nintendo's Mario franchise. The performance level in the game did not meet her/his expectation level. This tweet expresses a pain point as it clearly shows the problem of underperformance from the customer's perspective and offers actionable insight—improving the performance of the game.

Assessment performed by the NN model: 1 (Contains pain point)



```
@thetruebowser @nintendoamerica beautiful game but totally unacceptable performance. a huge let down coming from nintendo in te
rms of performance. please fix this nintendo.
Expresses pain point? 1
```

Figure 9: Automated detection of customer's pain points. Assessment—positive. The NN model has been able to make the correct assessment.

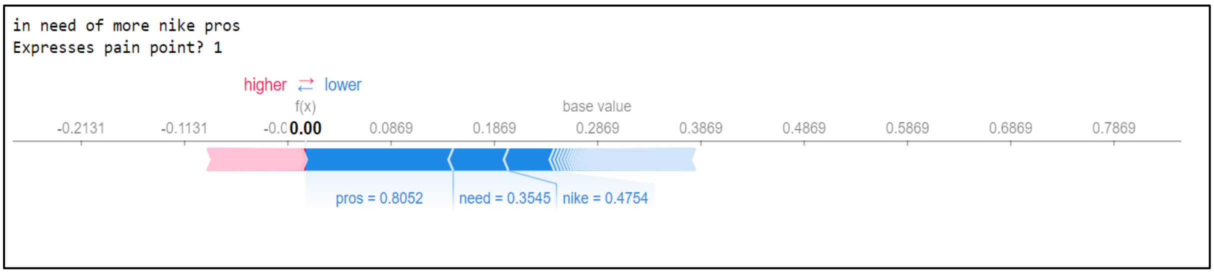
Final outcome: The NN model has been able to make the correct assessment.

Example 3:

Original tweet: "in need of more nike pros"

Our interpretation: In this tweet, the customer has stated her/his need for more Nike Pro products, a line of training apparel from Nike. However, the customer does not express any disappointment or psychological gap toward the company itself. Hence, the tweet does not contain any pain points.

Assessment performed by the NN model: 1 (Contains pain point)



```
in need of more nike pros
Expresses pain point? 1
```

Figure 10: Automated detection of customer's pain points. Assessment—positive. The NN model was not able to make the correct assessment.

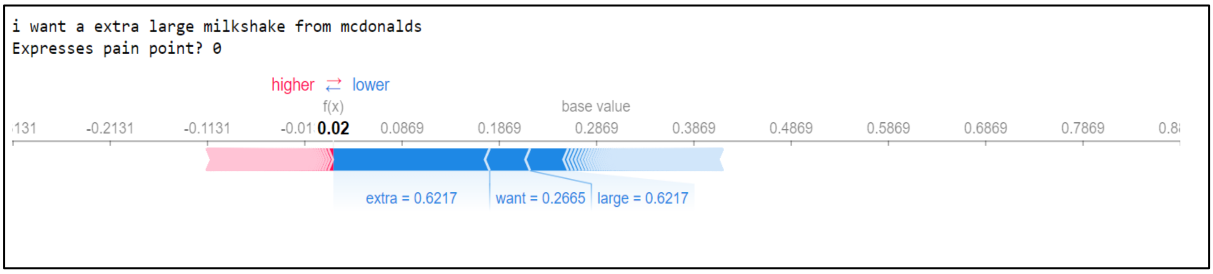
Final outcome: The algorithm has not made the correct assessment. The tweet consists of the words "need" along with "pros," which is often used colloquially to denote "problems." Hence, the algorithm calculated the value is "1" (i.e., the tweet expresses a pain point). We have deliberately selected the result of this particular tweet to present to the reader, as it is a clear demonstration that on some occasions, some tweets, or natural human language in general, can be confusing to the algorithms. Even though we have applied the latest advancements in ML to develop the NN model, it still lacks the higher-order comprehension that is typical of humans.

Example 4:

Original tweet: "I want a extra large milkshake from mcdonalds"

Our interpretation: The customer is simply expressing a desire in this tweet that she/he wants a particular product; a large-size milkshake from the fast-food chain McDonald's. Here, there is no expression of any form of dissatisfaction or unhappiness with the seller's offering. Hence, the tweet does not contain any pain points.

Assessment performed by the NN model: 0 (Does not contain pain point)



```
i want a extra large milkshake from mcdonalds
Expresses pain point? 0
```

Figure 11: Automated detection of customer's pain points. Assessment—negative. The NN model has been able to make the correct assessment.

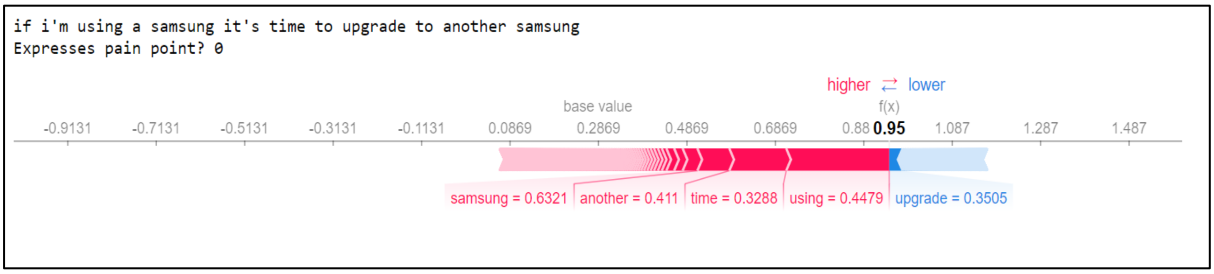
Final outcome: The algorithm has been able to make the correct assessment.

Example 5:

Original tweet: "if I'm using a Samsung it's time to upgrade to another Samsung"

Our interpretation: The customer is talking about using some type of consumer electronics from Samsung electronics, probably a mobile phone. In the tweet, he/she has expressed an opinion that if the customer is already using a piece of equipment from the seller, it was time to upgrade to another piece of equipment from the same firm. The tweet does not contain a pain point.

Assessment performed by the NN model: 0 (Does not contain pain point)



```
if i'm using a samsung it's time to upgrade to another samsung
Expresses pain point? 0
```

Figure 12: Automated detection of customer's pain points. Assessment—negative. The NN model has been able to make the correct assessment.

Final outcome: The algorithm has been able to make the correct assessment.

4.3 Feature Importance Analysis

For the interested reader, in this section, we offer further analysis on the inner workings of the NN model to automatically generate customer insights from user-generated content. In doing so, we analyze the contributions of both BERT features (for context-embeddings) and TF-IDF (that assigns scores to each word) (Horev, 2018; Salminen et al., 2019). However, complex ML models are not easy to interpret. Thus, we also deploy another recent improvement in ML—referred to as SHAP values—to analyze the performance. SHAP (**SH**apley **A**dditive ex**P**lanations) is a game-theoretic approach to explaining an ML model's performance. It uses traditional Shapley values from game theory and related extensions to link optimal credit allocation with local explanations (*SHAP Latest Documentation*, 2018).

Figure 13 shows that the BERT features have a considerably larger influence on the model's performance than the TF-IDF features, as the Top 20 features do not contain TF-IDF features at all. These findings indicate that bidirectionally trained language models have a better understanding than word-frequency-based representations of the meaning of language for paint-point detection, supporting similar findings in other NLP tasks (Devlin et al., 2019), and our findings imply these transformer-based feature representations also work the best for pain point detection.

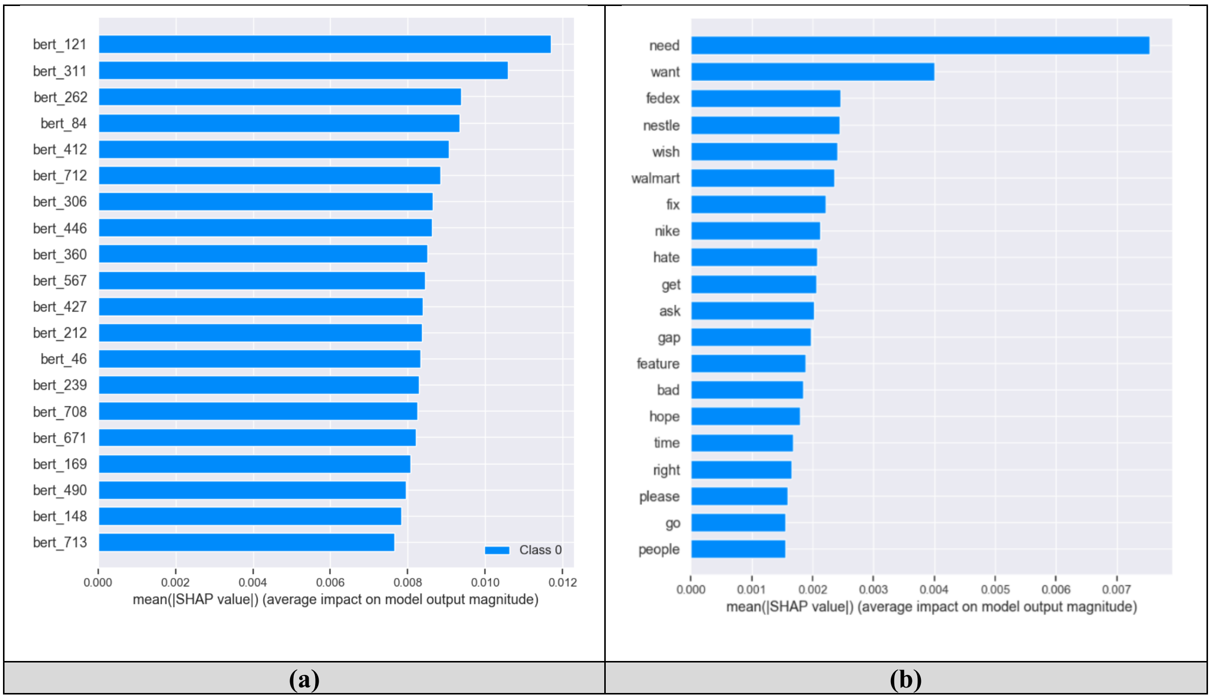


Figure 13: Influence of different features in the performance of the neural network. The TOP 20 highest features are all BERT features (a), whose drawback is that they are not human-interpretable (e.g., we cannot say what "bert_713" means since this is a numerical representation rather than a word). To identify impactful words, we ran another model interpretation analysis with TF-IDF features (b), which are interpretable by humans.

Important to note here is that even though the BERT features influence the performance of our NN model for automatic detection of customers' pain points, the individual BERT "Transformer" (bert-121, bert-311, etc., in Figure 13a) are not interpretable by humans. On the other hand, analyzing the TF-IDF features shows the importance of individual words in the model's overall performance, as we present in Figure 13b. This comparison allows understanding the words that have the most impact on the model, with words such as "need," "want," "wish," and "fix" ranking predominantly high for positive cases. The overall finding is that even though these specific indicator words provide some signal for pain point detection, transformer, such as BERT, are able to process in context in a way that substantially improves the prediction accuracy; in a word, pain point detection methods benefit from transformer models.

4.4 Additional Experiments

We sought ways to improve the classification performance based on the preliminary experiments. A two-staged plan was devised, which included (a) acquiring more training data, and (b) testing a more advanced state-of-the-art transformer for the text classification (i.e., RoBERTa (Y. Liu et al., 2019)). We hypothesize that increasing the dataset size would help the classifier to better separate pain point tweets from those not containing a pain point. This task would be better achieved with a classifier that has had robust performance in similar tasks (e.g., fake review detection (Salminen et al., 2022), which is a technically similar problem – i.e., text classification with binary classes.

First, to increase the dataset size, a sample of 2500 tweets were randomly selected among the corpus, out of those not previously included in the training data. These tweets were labeled following the same coding procedure as previously, i.e., the same researcher carrying out the coding and the same researchers validating the appropriateness of the coding. With the addition of these tweets, of which 592 (23.7%) were labeled to contain a pain point and 1908 (76.3%) to not contain a pain point, the total dataset size was now 4600 tweets[2], of which 1252 (27.2%) contained a pain point and 3348 (72.8%) did not contain a pain point (see Figure 14).

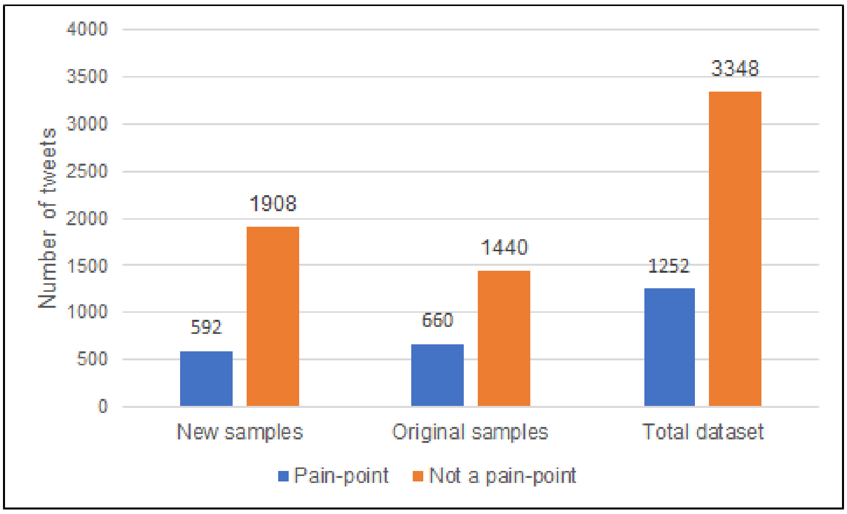


Figure 14: Second round data collection, original dataset, and the combined total dataset.

[2] In this number, we also included 100 tweets that had been cut out from the first version of the dataset, i.e., FedEx had 100 more tweets than the rest and we wanted to have an exact brand balance in the original dataset.

To carry out the experiments, preprocessing was done on individual tweets to remove hashtags (#), mentions (@), and URLs. As most sentences after cleaning laid between 30-40 length, 35 was chosen as the maximum length ('MAX LENGTH') parameter to be fed to transformer models during training. Otherwise, the majority of the sentences would have been padded with 0 till the allowed maximum length, which would have been of no meaning for the models. After experimenting with different combinations of parameters for batch size ('BATCHSIZE') and learning rate ('LR'), two important parameters for deep learning (Smith, 2018), the best performance was obtained with the RoBERTa classifier using the following parameters: length=35, batch size=16, and learning rate=1e-5. (The details of the experiments are available in computational notebooks provided as supplementary material: **[link hidden for anonymous review])**.

Table 4: RoBERTa results on the expanded pain point dataset (the numbers in parentheses are from the best performing BERT model, bert-base-uncased).

Class	**Precision**	**Recall**	**F1-score**	**Support**
No pain point	0.87 (0.85)	0.94 (0.90)	0.90 (0.88)	332 (332)
Has pain point	0.80 (0.70	0.62 (0.59)	0.70 (0.64)	128 (128)
Accuracy	0.85 (0.82)	n/a	n/a	460 (460)
F1 macro	0.83 (0.78)	0.78 (0.74)	0.80 (0.76)	460 (460)
F1 weighted	0.85 (0.81)	0.85 (0.82)	0.85 (0.81)	460 (460)

The results, including the best BERT models (two were tested: *bert-base-uncased* and *bert-base-cased*, are shown in Table 4. Because the dataset was imbalanced, the most important metric is the macro-F1 score, i.e., the average of the harmonic means of precision (positive predictive value) and recall (true positive rate) for the pain point and non-pain point classes. The best RoBERTa results (macro-F1 = 0.80 and accuracy 85%) are a considerable improvement from the preliminary experiments, where the macro-F1 score was 0.66 and accuracy was 76.5%. Most importantly, the new results indicate a sizeable improvement over a random guess. If we assume a non-weighted random guess of 50% (i.e., an equal chance of a random classifier to state a given tweet is or is not a pain point), the improvement over this baseline is (80-50) / 50 = 60%. Given the class imbalance, we can consider a case where the random classifier would always predict a tweet is not a pain point (as 73% of the tweets belong to this class). By always guessing “not a pain point,” the classifier could achieve the maximal accuracy of 73%, i.e., it would correctly classify all the non-pain point cases and misclassify all real pain points. Relative to this probability, our RoBERTa classifier achieves an improvement of (80-73) / 73 = 10 %

This is not a trivial improvement but more important is that such a baseline would detect *none* of the real pain point tweets, so it would be useless in practice. Also, its macro-F1 would be capped to 0.50 because only one class would be correctly predicted. This means that the improvement over macro-F1 would be the same as a true random guess, i.e., 60%.

Overall, these results can be considered satisfactory given the somewhat difficult nature of automatic pain point detection (with the challenges mentioned in Section 2.3).

4.5 Pain point Types

In addition, we wanted to probe deeper into the types of customer pain points. This was achieved by inductively investigating the samples with a positive class (i.e., the tweets that the best classifier considered as pain points) among the manually annotated samples (N=1252 pain points), and generating a descriptive taxonomy of pain point types. This taxonomy ended up having five categories, which are shown in Table 5. These categories were considered in a multiclass classification task, whose purpose was to classify pain point tweets into one of the five classes (i.e., pain point types).

Table 5: Pain point categories inductively formulated from the pain point tweets. 'Count' indicates the number of annotated cases.

Type of Pain	Count	Definition	Examples
Operational issues	N=356 (28.4%)	A range of pain points that are linked to the companies' various types of operational issues, for example, sales, technology-used, logistics, communication, or quality of business premises.	Examples include product unavailability, unfavorable operating hours, non-functioning payment terminals, packaging and tracking issues in case of online purchasing, quality of business premises: Dirty and unhealthy dining area.
Product feature or quality	N=559 (44.6%)	The customers express negative feelings due to various issues that can be attributed to the quality or features of the products.	Examples include the quality of food at McDonald's, the product-range offered by a retailer, or security issues of online products. Complaints regarding product pricing are included in this category, too.
Company's image	N=96 (7.7%)	The customers express uneasiness about conducting business with a company due to issues that portray a negative image of the company. The customers consider that these issues can be solved by the company.	Use of child labor; Treating employees in an unfair manner.
Customer service	N=89 (7.1%)	The customers express negative experiences with the various types of customer service that they received or sought for, both on and offline.	Examples include both face-to-face interactions with the employees, for example, unhelpful personnel at retailer's checkout point, and online interactions, for instance, perceived rude customer service agent over the phone, or complex online fraudulent claim systems.
Service quality or failure	N=152 (12.1%)	Customers express disappointments as their perceived service quality was inferior to expected service quality	Discounts for new customers that are perceived unfair by existing loyal customers; customers having negative experience due to other customers.

This dataset poses an even more challenging task for classification than the complete dataset, namely because (a) *it only contains the subset of positive cases* (i.e., the pain point tweets, N=1252), and (b) *the class imbalance problem is now targeting five classes as opposed to one* (see Figure 15). Therefore, we tackle these two issues via a strategy based on three steps that involve experimenting with (a) balanced sample sizes, (b) RoBERTa against BERT, and (c) various data augmentation techniques. The first step involved testing with 50, 100, and 200 samples randomly drawn from each class (unless the number of samples in the class was smaller, in which case all the samples were selected).

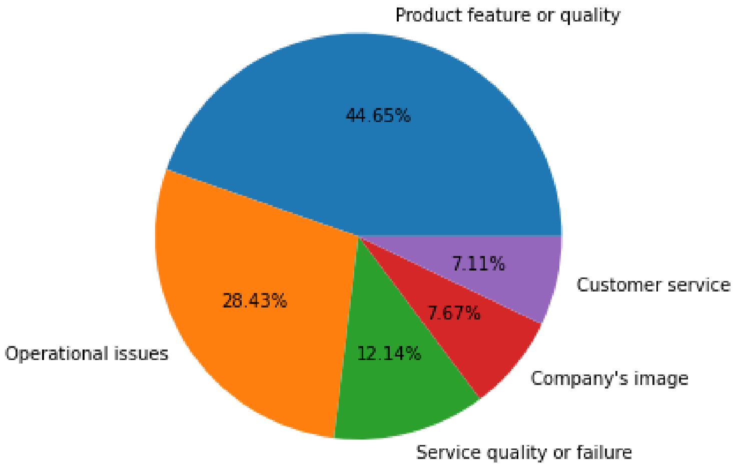


Figure 15: Distribution of pain point type samples.

We created a classification layer over the output of the transformer model output (BERT/RoBERTa) so we could leverage the pre-trained transformer models without retraining them from zero. In other words, we are using the pretrained models to perform a downstream task of text classification with five different categories. We used 75/25 train-test split for all experiments. These results showed that *100 samples* produced the best overall result (macro-F1 = 0.54) relative to 50 (macro-F1 = 0.42) and 200 (F1 = 0.49). More specifically, 50 was too little for a signal on any of the classes, whereas 200 biased the result excessively in favor of the dominant class (‘Product feature or quality’). The experiments comparing transformers showed that RoBERTa outperformed BERT by approximately 6% in terms of F1 score – this was important to test because the problem type was different from previously (multiclass instead of binary classification).

Finally, we tested five data augmentation tactics: DELETION, SWAPPING, SYNONYM, SUBSTITUTE (distil-BERT), and SUBSTITUTE (FastText). These tactics were implemented using the NLPaug library[3]. **DELETION**, also known as Random Word Deletion (RWD) (Arras et al., 2017), duplicates the training set sentences and then randomly deletes a word from each sentence. **SWAPPING** duplicates the dataset and randomly changes the order of words in the sentences to introduce variability (Y. Zhang et al., 2019). **SYNONYM** uses Princeton University's WordNet lexicon that contains semantic relations between words (Miller, 1995); here, we locate the closest synonym to a randomly selected word in a sentence and replace it with this word in the augmented dataset. **SUBSTITUTE** uses a pre-trained word embedding model to create a sentence that has the same meaning but uses different words; here, we test with two models, *distil-BERT* (Sanh et al., 2019) and *FastText* (Joulin et al., 2016) because of their efficiency and robust performance in NLP tasks.

The results showed that the best data augmentation tactic was SYNONYM augmentation. We repeated the tests in total five times, and although the DELETE model achieved the highest single-time performance, its performance declined rapidly in repeated runs. This indicates that the high performance score it obtained was a singular case rather than the result of true ability to correctly classify the samples. In turn, the SYNONYM model showed consistently high performance, and as its highest obtained score was the second-highest among the tested augmentation models, we report this model as the best pain point type detector and applied it when classifying the pain point tweets in the wild.

[3] https://nlpaug.readthedocs.io/en/latest/augmenter/word/word_embs.html

Based on these experimental findings, the best model was RoBERTa 100 samples using SYNONYM augmentation, clearly outperforming RoBERTa 100 samples without any data augmentation (see Table 6).

Table 6: Pain point type classification results.

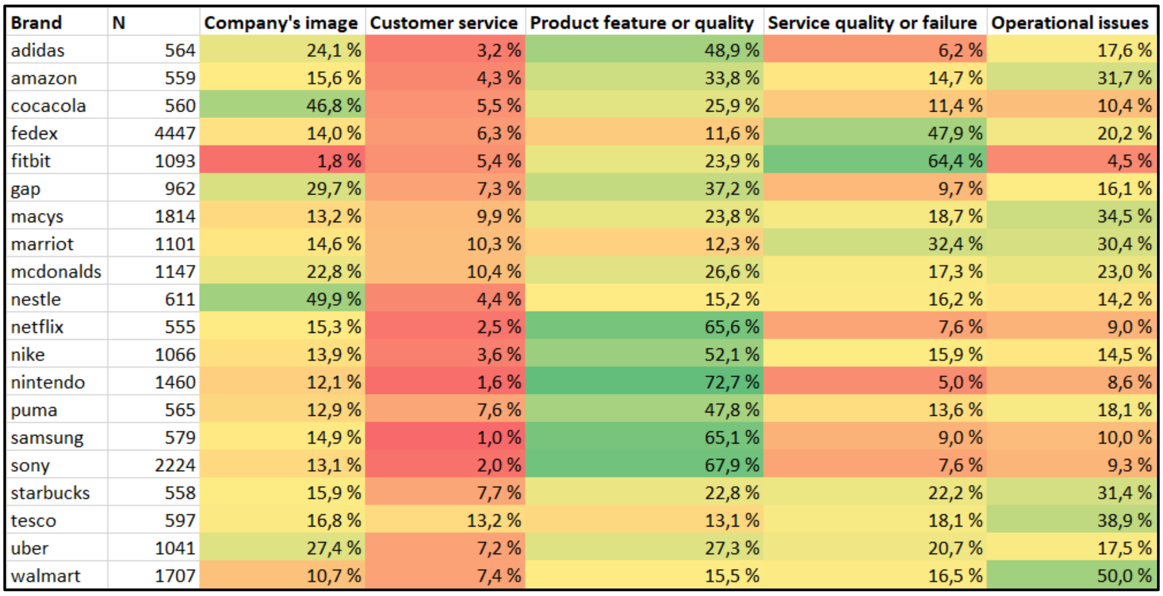
	Company's image	**Customer service**	**Operational issues**	**Product feature or quality**	**Service quality or failure**	**Weighted**
RoBERTa (*no* augmentation)	0.67	0.56	0.39	0.71	0.64	0.61
RoBERTa (*with* SYNONYM augmentation)	0.79	0.71	0.68	0.85	0.64	0.74
Improvement from augmentation	+17.9%	+26.8%	+74.4%	+19.7	+0%	+21.3%

4.6 Pain point Profiling

We now demonstrate how to apply the two trained models – i.e., the pain point detector and the pain point type detector – for pain point profiling, which we define as determining the prevalence of different pain point types that the firm has towards deriving "managerial insights to make better decisions and improve performance" (Verhoef et al., 2010), which is considered a general goal of customer insights.

First, the best RoBERTa pain point detection model was used for classifying the full dataset of 4.2M tweets – among these, the model found 23,210 pain point tweets (0.55% of total). The small percentage indicates that either the prevalence of pain points in the wild is lower than in the training data or the model has high precision. Further experiments are required for testing these suppositions, which we leave for future work—here, we proceed with the pain point type detection using the 23K dataset of pain points in the wild.

The best RoBERTa pain point type detection model was applied to this dataset, thus obtaining the frequency of pain point tweets by brand and type. Since each brand had a varying number of pain point tweets, we show the pain point profiles as ratios (see Figure 16).



Brand	N	Company's image	Customer service	Product feature or quality	Service quality or failure	Operational issues
adidas	564	24,1 %	3,2 %	48,9 %	6,2 %	17,6 %
amazon	559	15,6 %	4,3 %	33,8 %	14,7 %	31,7 %
cocacola	560	46,8 %	5,5 %	25,9 %	11,4 %	10,4 %
fedex	4447	14,0 %	6,3 %	11,6 %	47,9 %	20,2 %
fitbit	1093	1,8 %	5,4 %	23,9 %	64,4 %	4,5 %
gap	962	29,7 %	7,3 %	37,2 %	9,7 %	16,1 %
macys	1814	13,2 %	9,9 %	23,8 %	18,7 %	34,5 %
marriot	1101	14,6 %	10,3 %	12,3 %	32,4 %	30,4 %
mcdonalds	1147	22,8 %	10,4 %	26,6 %	17,3 %	23,0 %
nestle	611	49,9 %	4,4 %	15,2 %	16,2 %	14,2 %
netflix	555	15,3 %	2,5 %	65,6 %	7,6 %	9,0 %
nike	1066	13,9 %	3,6 %	52,1 %	15,9 %	14,5 %
nintendo	1460	12,1 %	1,6 %	72,7 %	5,0 %	8,6 %
puma	565	12,9 %	7,6 %	47,8 %	13,6 %	18,1 %
samsung	579	14,9 %	1,0 %	65,1 %	9,0 %	10,0 %
sony	2224	13,1 %	2,0 %	67,9 %	7,6 %	9,3 %
starbucks	558	15,9 %	7,7 %	22,8 %	22,2 %	31,4 %
tesco	597	16,8 %	13,2 %	13,1 %	18,1 %	38,9 %
uber	1041	27,4 %	7,2 %	27,3 %	20,7 %	17,5 %
walmart	1707	10,7 %	7,4 %	15,5 %	16,5 %	50,0 %

Figure 16: Pain point profiling based on the 4M tweets analyzed. The percentages indicate how a brand's pain point tweets are distributed along with the five identified pain point categories.

The pain point profiles in Figure 16 can be read in two ways: to identify concerns (a) per brand and (b) per pain point type. For example, we can observe that for Adidas, the biggest issues are related to product features and quality (48.9% of the pain point tweets). This category seems particularly strong for other companies as well, including Netflix (65.6%), Nike (52.1%), Nintendo (72.7%), Puma (47.8%), Samsung (65.1%), and Sony (67.9%). Apart from Netflix, all of these are manufacturing companies, so the prevalence of this category makes intuitive sense. By exploring the type of complaints within this category, the brands can obtain ideas to improve their product quality and features. Despite the prominence of the product feature and quality category, it is not the largest category for all brands. For example, company image issues are the biggest concern for Coca-Cola (46.8%) and Nestlé (49.9%). Relative to other brands, Tesco has the most issues with customer service (13.2%). Users most report service quality or failure regarding FedEx (47.9%) and Fitbit (64.4%), whereas operational issues are most logged for Walmart (50.0%), Tesco (38.9%), Starbucks (31.4%), Macy's (34.5%), Marriott (30.4%), and Amazon (31.7%).

5 DISCUSSION AND IMPLICATIONS

5.1 General Discussion

Generating customer insights is a continual challenge facing companies across a broad range of industries and markets (Price & Wrigley, 2016). Two core challenges are access to a sufficient amount of data, the proper analytical tools, and procedures to analyze them to generate customer insights (Handfield & Steininger, 2005; B. Wang et al., 2016). The application of ML technologies to collect and analyze the publicly available UGC can help to overcome both the challenges to a large extent. In this study, we develop and demonstrate how ML models can be developed and trained to automatically identify customers' pain points and the type of these pain points in order to generate customer insights.

Identifying pain points helps to generate insights on the specific problems that customers experience with firms and their offerings that the companies can address (Homburg & Fürst, 2007; Rawson et al., 2013; B. Wang et al., 2016). However, their diverse and varied nature, along with the immense challenge of analyzing an enormous amount of unstructured data, make their detection particularly challenging (Abu-Salih et al., 2018; Salminen et al., 2019). Our findings show that a neural network using transformer features offers the best performance in terms of pain point and pain point type detection.

Because of the greatly increased availability, sophistication, and relevance of data, there has been a growing shift in the field of marketing from traditional forms of content analysis to more sophisticated computational forms (Balducci & Marinova, 2018; Kumar, 2018; X. Liu et al., 2016). Simultaneously, there has been a call for a parallel advancement of marketing research methodology (Davis et al., 2013; Hofacker, 2012), so new approaches can help advance marketing theory, especially by extracting deeper insights from unstructured, multifaceted, and non-linear data (Davis et al., 2013; Syam & Sharma, 2018). We contribute to this goal by developing and demonstrating a method for taking unstructured online material, cleaning and structuring it, and training the most modern and capable ML algorithms to generate customer insights (Homburg & Fürst, 2007; B. Wang et al., 2016).

UGC in social media reveals that consumer perceptions of companies are multifaceted (J. Zhang et al., 2021), of which pain point detection is undoubtedly one of these areas of the voice of the consumer. This research shows that pain point detection via machine learning is a promising area that deserves further research. The web can be seen as a platform for customer insights towards enhanced product innovation and process improvements (Sawhney et al., 2005) — in other words, analyzing the tweets can reveal cases of "what went wrong" towards ideas for product development and design. Therefore, the utility of detecting pain points is not constrained to marketing use cases but deals with other areas as well, such as product development, design, and engineering, thereby highlighting the role of the marketing function in the organization as a curator of customer insights and acting as a middleman between product development and customer needs.

5.2 Research Contribution

As the literature review illustrated (see Section 2.2), preliminary work on customer needs detection, needmining, needfinding, and pain point detection have all conceptually similar problems. However, these studies tend to experiment with a limited number of algorithms, often do not interpret the model's decisions, typically do not divide customer needs into sub-types, and rarely make the datasets and/or models publicly available for other researchers. Therefore, the problem of pain point detection is still very much active and unresolved by the academic marketing community. Our study addresses some of the said shortcomings, e.g., by experimenting with several algorithms, providing example predictions of pain points as well as feature analysis, considering not only pain point detection but also pain point type detection, and by making the models and the dataset publicly available for others. Through these contributions, we hope to encourage subsequent work that continues to address the grand challenge of inferring customer insights from the voice of the customers.

We also contribute by generating understanding about pain point detection as a problem that involves some characteristic traits: (1) *class imbalance*, i.e., pain points are rarer in the wild than other types of brand mentions (this finding is also supported by Kühl et al. (2020)), and (2) *ambiguity* – although this trait, again, is supported by previous research (Kühl et al., 2019), we want to highlight it because we went through several iterations of definition-example giving-annotating-reliability check, and still were not able to achieve perfect agreement among the expert coders. So, achieving high-quality data using a resource like anonymous crowd workers would be highly risky, in our opinion. Instead, we believe that the better is to accept that the problem entails a degree of inherent subjectivity (Alonso, 2015) and work among the researchers to achieve a consensus on the ambiguous before deciding their final label.

These traits also affect what kind of algorithms and computational approaches (e.g., pre-processing, data augmentation) could and should be experimented with – therefore, this characterization is useful for others working in this problem domain.

5.3 Managerial Contributions and Recommendations

From a managerial perspective, a pain point is something that a customer is aware of and is bothered by. It is a problem waiting for a solution. For a firm, the first step toward the solution is identifying those pain points appropriately. Our study can be beneficial for companies toward that goal. Although modern data-driven business scenarios can often be characterized by the abundance of large volumes of data (Kumar, 2018), preparing the data and structuring it to be of actionable value to a business is challenging (Syam & Sharma, 2018). We offer concrete and applicable solutions to a highly relevant problem within these contextual challenges.

Moreover, managers can grasp the core notion of our study and modify the application in such manners that fulfill their particular needs. Businesses and entrepreneurs may craft apt value propositions that draw customers to solve their problems by finding pain points. These can be product pain points (for example, faulty or non-functional products), efficiency pain points (for example, the difficulty of using the product), financial pain points (for example, high repeat purchasing costs), and process pain points (for example, delayed support response) are all common examples (Patel, 2020; B. Wang et al., 2016).

Discovering customer pain points impacts both a firm's sales and its marketing strategy. The sales team identifies the pain points to tailor their pitch and present the products as the right solution (Patel, 2020; B. Wang et al., 2016). Marketers want to understand these pain points so that they can advertise their solution effectively in an appealing way. After identifying pain points, companies can figure out key solutions for them, which in turn can enhance product attractiveness and improve customer satisfaction effectively. Moreover, identifying and eliminating pain points is essential in improving customer experience (Patel, 2020; B. Wang et al., 2016). One practical approach can be value co-creation through customer participation (Mustak et al., 2013), which can respond to customer concerns and thus address the communication gap between the brand and its customers. This aspect entails viewing the customer complaints as problems and a resource for creating a collaborative community (Weinberg et al., 2013).

Furthermore, our approach can help to curate and seed content by desired criteria (e.g., customer interests), which is beneficial for firms in content marketing or even formulating their business strategy. To that end, we recommend that organizations that use ML to track and control their performance regularly retrain the models when they fall below a certain threshold. This threshold value is domain-specific; for example, no universal F1 score value applies to all domains. Instead, businesses must develop and adhere to their standards for evaluating the quality of the models.

From a technical perspective, the performance of the models varies by brand, which implies that to get the optimal result, companies should be advised to generate their own dedicated classifiers, which can be done at a reasonable cost using modern ML and NLP technologies and is, as our results have shown, feasible in terms of accuracy. Most customer-generated comments targeted to a brand do not contain a pain point, which implies that the algorithms best suited for this task are those that can manage class imbalance. The length of the user's comments seems to be associated with the probability of containing a pain point, which implies that heuristics can provide some degree of signal for pain point detection (albeit, in our tests, the best heuristic model was the one using keywords and sentiment labels).

Alongside identifying improvement points for their internal use, the brands can apply pain point profiling for competitive analysis (Deshpande & Gatingon, 1994), i.e., by comparing their pain point profile against that of competitors. Such an analysis helps discover how the brand is positioned in consumers' minds, i.e., that it often seems to have product quality issues or a relatively higher share of operational issues.

While there is a general strive for automating downstream decision-making tasks based on customer insights (e.g., "[there are] calls for substantial research to develop statistical algorithms that measure customer insights and to develop optimization routines as decision-support systems to automate the implementations of marketing decisions for better management of customer relationships." (Sun et al., 2006, p. 19), we recommend that practitioners apply caution with fully automated decision-making processes. This is due to two reasons: (1) first, the performance of any ML model is not perfect, which means there are both false positives and false negatives among the results, even though the general accuracy would be high. Therefore, while the model performs well for the use case of aggregate customer analysis, it is more precarious for the use of individual customer cases. Second, (2) the nature of pain points requires comprehension and human analysis to truly draw useful insights from them: whether product development ideas, service process improvement, or other actionable information, these plans require holistic thinking and understanding of the world that a ML model simply does not have. Therefore, the role of the model in marketing decision making is to summarize and aggregate information for an overall analysis and, on the other hand, provide specific example cases for a human marketer to analyze in depth. While human sense-making of the content is a crucial step (Rydén et al., 2015), the value from the latter use case is still substantial because pre-filtering pain point types from the enormous number of tweets on the wild will save marketers' time.

Finally, the models could be implemented towards developing "social CRM" systems (Malthouse et al., 2013), i.e., social media monitoring systems that inform decision makers working for brands about the 'pain point sentiment' of their online audiences. Because tweets contain a timestamp (i.e., time of creation), plotting how a specific pain point type for a given brand evolves is possible. By deploying trend detection algorithms (Kämpf et al., 2015), one could create a system that alerts the marketing decision maker when a given pain point type is "spiking," i.e., is anomalously high. Early detection of such spikes could offer firms ways to proactively address evolving concerns, which is an area that currently has limited viable solutions (Plangger & Montecchi, 2020).

5.3 Limitations and Future Research Directions

Our study is restricted by limitations, which also opens up the possibility for further research in this promising area. Indeed, obtaining more data and training the algorithms on their basis to develop the models' capabilities would be one addition to our research. For our further experiments, sentiment analysis heuristic methods were promising in terms of performance improvement. There are several possible underlying reasons for this ranging from sentiment, beyond a few heuristic-based terms, is noisy to the underlying dataset to algorithmic performance of difference sentiment algorithms. So, given the prevalence of sentiment analysis in marketing research, this is certainly a fruitful area for future research. Moreover, we have trained our ML models based on tweets in the English language only. Future studies that may push this language barrier and expand the application to other languages will be valuable.

Our study is built upon a Twitter-based dataset. We recommend future research to extend the applicability of our approach through utilizing other forms of unstructured data, for example, customers' review platforms, especially in the UGC content area, of which pain points are one. Given the promising results presented in this research, it is clear that pain points is an area warranting more research. This expansion may even include other forms of online content, such as images and videos, where ML has also shown high promise (Salminen et al., 2019). Given the diversity of text-based data sources, data integration is a key challenge for brands (Verhoef et al., 2010). Customers address the firm on Twitter and other social media channels (e.g., Facebook, Instagram, SnapChat, etc.) and via private channels (email, WhatsApp, Messenger, etc.). We focused on Twitter because tweets are publicly available, unlike posts in most other channels. Nevertheless, integrating pain points from the entire channel mix remains an avenue for future development and research. While data integration poses challenges, a helpful aspect is that our approach can be applied to any text, and therefore, our approach supports both online and offline data, any source where customer writings about a brand can be extracted. One emerging opportunity is *speech*, i.e., transcribing customers' encounters with call agents (Argyris et al., 2021) and analyzing these transcripts.

We have investigated data (tweets) that are specifically targeted towards specific firms and their offerings. Future studies that examine "non-targeted" or "open" complaints may offer deep insights into customers' untapped needs and create the possibility to develop new value propositions or even business models.

Further segmentation of the Twitter users could be done to identify candidate lead users – lead users are those that currently have needs that will become common in the marketplace only in the future (Von Hippel, 1986). By identifying these lead user needs and addressing them early on, a firm can achieve a competitive advantage (Urban & Von Hippel, 1988; Von Hippel, 2009). Our current approach considers each user's "voice" of equal importance, but by applying the lead user theory, some concerns would be more revelatory and strategically important than others. This is an area for future research.

Considering the replicability of the results, a fixed seed was kept for training, validation, and test dataset splitting, while seeds were set using torch implementation to keep training result consistent across multiple runs of the training algorithm. We share the best trained models, along with computational notebooks, and the *Pain point dataset* that includes both the pain point and the pain point type annotations, in a GitHub repository: [**link hidden for anonymous review**]. These contributions enable firms and researchers to further develop pain point detection algorithms and systems.

6 CONCLUSION

Leveraging the benefits of ML applications in marketing and addressing the important need for such applications for marketing research methods, this research contributes to the literature by demonstrating how to identify customers' pain points through supervised training of ML algorithms. Besides, we conducted a comparative analysis of the performance of six different ML models. We found that the neural network using transformer features performs the best for identifying pain points from customers' natural (colloquial) language-based expressions on Twitter. The model provides an accuracy that is close to a human-to-human agreement and holds high promise for both marketing academia and practice.

REFERENCES

Abu-Salih, B., Wongthongtham, P., & Chan, K. Y. (2018). Twitter mining for ontology-based domain discovery incorporating machine learning. *Journal of Knowledge Management*, *22*(5), 949–981. https://doi.org/10.1108/JKM-11-2016-0489

Agrawal, T. (2020). *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient* (1st ed. edition). Apress.

Alonso, O. (2015). Practical Lessons for Gathering Quality Labels at Scale. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1089–1092. https://doi.org/10.1145/2766462.2776778

Alonso, O., Marshall, C. C., & Najork, M. (2015). Debugging a Crowdsourced Task with Low Inter-Rater Agreement. *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, 101–110. https://doi.org/10.1145/2756406.2757741

Amado, A., Cortez, P., Rita, P., & Moro, S. (2018). Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics*, *24*(1), 1–7. https://doi.org/10.1016/j.iedeen.2017.06.002

Antons, D., & Breidbach, C. F. (2018). Big Data, Big Insights? Advancing Service Innovation and Design With Machine Learning. *Journal of Service Research*, *21*(1), 17–39. https://doi.org/10.1177/1094670517738373

Argyris, Y. A., Monu, K., Kim, Y., Zhou, Y., Wang, Z., & Yin, Z. (2021). Using Speech Acts to Elicit Positive Emotions for Complainants on Social Media. *Journal of Interactive Marketing*, *55*, 67–80.

Arras, L., Montavon, G., Müller, K.-R., & Samek, W. (2017). Explaining recurrent neural network predictions in sentiment analysis. *ArXiv Preprint ArXiv:1706.07206*.

Asiri, S. (2018, June 11). *Machine Learning Classifiers*. Medium. https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, *46*(4), 557–590.

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, *84*(1), 1–25.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brownlee, J. (2016, August 16). A Gentle Introduction to XGBoost for Applied Machine Learning. *Machine Learning Mastery*. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

Chen, R., Wang, Q., & Xu, W. (2019). Mining user requirements to facilitate mobile app quality upgrades with big data. *Electronic Commerce Research and Applications*, *38*, 100889. https://doi.org/10.1016/j.elerap.2019.100889

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cheng, L.-C., Chen, K., Lee, M.-C., & Li, K.-M. (2021). User-Defined SWOT analysis – A change mining perspective on user-generated content. *Information Processing & Management*, *58*(5), 102613. https://doi.org/10.1016/j.ipm.2021.102613

Cui, D., & Curry, D. (2005). Prediction in Marketing Using the Support Vector Machine. *Marketing Science*, *24*(4), 595–615.

Cui, G., Wong, M. L., & Lui, H.-K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science*, *52*(4), 597–612.

Davis, D. F., Golicic, S. L., Boerstler, C. N., Choi, S., & Oh, H. (2013). Does marketing research suffer from methods myopia? *Journal of Business Research*, *66*(9), 1245–1250. https://doi.org/10.1016/j.jbusres.2012.02.020

Deshpande, R., & Gatingon, H. (1994). Competitive analysis. *Marketing Letters*, *5*(3), 271–287.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Donges, N. (2019). *A complete guide to the random forest algorithm*. Built In. https://builtin.com/data-science/random-forest-algorithm

Freelon, D. (2014). On the interpretation of digital trace data in communication and social computing research. *Journal of Broadcasting & Electronic Media*, *58*(1), 59–75.

Griffin, A., & Hauser, J. R. (1993). The Voice of the Customer. *Marketing Science*, *12*(1), 1–27. https://doi.org/10.1287/mksc.12.1.1

Gupta, S., Leszkiewicz, A., Kumar, V., Bijmolt, T., & Potapov, D. (2020). Digital analytics: Modeling for insights and new methods. *Journal of Interactive Marketing*, *51*, 26–43.

Handfield, R. B., & Steininger, W. (2005). An Assessment of Manufacturing Customer Pain Points: Challenges for Researchers. *Supply Chain Forum: An International Journal*, *6*(2), 6–15. https://doi.org/10.1080/16258312.2005.11517143

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, *36*(1), 20–38. https://doi.org/10.1016/j.ijresmar.2018.09.009

Hofacker, C. F. (2012). On Research Methods in Interactive Marketing. *Journal of Interactive Marketing*, *26*(1), 1–3. https://doi.org/10.1016/j.intmar.2011.10.001

Homburg, C., & Fürst, A. (2007). See no evil, hear no evil, speak no evil: A study of defensive organizational behavior towards customer complaints. *Journal of the Academy of Marketing Science*, *35*(4), 523–536.

Horev, R. (2018, November 17). *BERT Explained: State of the art language model for NLP*. Medium. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Humphreys, A., & Wang, R. J.-H. (2018). Automated Text Analysis for Consumer Research. *Journal of Consumer Research*, *44*(6), 1274–1306. https://doi.org/10.1093/jcr/ucx104

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1).

Johnson, D. S. (2007). Achieving customer value from electronic channels through identity commitment, calculative commitment, and trust in technology. *Journal of Interactive Marketing*, *21*(4), 2–22.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *ArXiv Preprint ArXiv:1612.03651*.

Kämpf, M., Tessenow, E., Kenett, D. Y., & Kantelhardt, J. W. (2015). The Detection of Emerging Trends Using Wikipedia Traffic Data and Context Networks. *PLOS ONE*, *10*(12), e0141892. https://doi.org/10.1371/journal.pone.0141892

Kingma, D. P., & Ba, L. J. (2015). *Adam: A Method for Stochastic Optimization*. 3rd International Conference for Learning Representations. https://dare.uva.nl/search?identifier=a20791d3-1aff-464a-8544-268383c33a75

Klapdor, S., Anderl, E. M., von Wangenheim, F., & Schumann, J. H. (2014). Finding the Right Words: The Influence of Keyword Characteristics on Performance of Paid Search Campaigns. *Journal of Interactive Marketing*, *28*(4), 285–301. https://doi.org/10.1016/j.intmar.2014.07.001

Knocklein, O. (2019, June 15). *Classification Using Neural Networks*. Medium. https://towardsdatascience.com/classification-using-neural-networks-b8e98f3a904f

Kohli, A. K., & Jaworski, B. J. (1990). Market orientation: The construct, research propositions, and managerial implications. *The Journal of Marketing*, *54*(2), 1–18.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*(1), 3–24.

Kramer, O. (2013). K-nearest neighbors. In *Dimensionality reduction with unsupervised nearest neighbors* (pp. 13–23). Springer.

Kranzbühler, A.-M., Kleijnen, M. H. P., & Verlegh, P. W. J. (2019). Outsourcing the pain, keeping the pleasure: Effects of outsourced touchpoints in the customer journey. *Journal of the Academy of Marketing Science*, *47*(2), 308–327. https://doi.org/10.1007/s11747-018-0594-5

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems*, 231–238.

Kühl, N., Mühlthaler, M., & Goutier, M. (2019). Supporting customer-oriented marketing with artificial intelligence: Automatically quantifying customer needs from social media. *Electronic Markets*, *30*(2), 351–367. https://doi.org/10.1007/s12525-019-00351-0

Kühl, N., Scheurenbrand, J., & Satzger, G. (2020). Needmining: Identifying micro blog data containing customer needs. *ArXiv:2003.05917 [Cs]*. http://arxiv.org/abs/2003.05917

Kumar, V. (2018). Transformative Marketing: The Next 20 Years. *Journal of Marketing*, *82*(4), 1–12. https://doi.org/10.1509/jm.82.41

Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research*, *48*(5), 881–894. https://doi.org/10.1509/jmkr.48.5.881

Li, H., Chen, Q., Zhong, Z., Gong, R., & Han, G. (2022). E-word of mouth sentiment analysis for user behavior studies. *Information Processing & Management*, *59*(1), 102784. https://doi.org/10.1016/j.ipm.2021.102784

Li, Y., & Yuan, Y. (2017). Convergence Analysis of Two-layer Neural Networks with ReLU Activation. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/a96b65a721e561e1e3de768ac819ffbb-Abstract.html

Libai, B., Bart, Y., Gensler, S., Hofacker, C. F., Kaplan, A., Kötterheinrich, K., & Kroll, E. B. (2020). Brave new world? On AI and the management of customer relationships. *Journal of Interactive Marketing*, *51*, 44–56.

Liu, L., Dzyabura, D., & Mizik, N. (2020). Visual Listening In: Extracting Brand Image Portrayed on Social Media. *Marketing Science*, *39*(4), 669–686. https://doi.org/10.1287/mksc.2020.1226

Liu, X., Burns, A. C., & Hou, Y. (2017). An investigation of brand-related user-generated content on Twitter. *Journal of Advertising*, *46*(2), 236–247.

Liu, X., Singh, P. V., & Srinivasan, K. (2016). A structured analysis of unstructured big data by leveraging cloud computing. *Marketing Science*, *35*(3), 363–388.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv:1907.11692 [Cs]*. http://arxiv.org/abs/1907.11692

Liu, Y., Soroka, A., Han, L., Jian, J., & Tang, M. (2020). Cloud-based big data analytics for customer insight-driven design innovation in SMEs. *International Journal of Information Management*, *51*, 102034. https://doi.org/10.1016/j.ijinfomgt.2019.11.002

Ma, L., & Sun, B. (2020). Machine learning and AI in marketing – Connecting computing power to human insights. *International Journal of Research in Marketing*, *37*(3), 481–504. https://doi.org/10.1016/j.ijresmar.2020.04.005

Macdonald, E. K., Wilson, H. N., & Konuş, U. (2012, September 1). Better Customer Insight—In Real Time. *Harvard Business Review*. https://hbr.org/2012/09/better-customer-insight-in-real-time

Malthouse, E. C., Haenlein, M., Skiera, B., Wege, E., & Zhang, M. (2013). Managing customer relationships in the social media era: Introducing the social CRM house. *Journal of Interactive Marketing*, *27*(4), 270–280.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Mustak, M., Jaakkola, E., & Halinen, A. (2013). Customer participation and value creation: A systematic review and research implications. *Managing Service Quality: An International Journal*, *23*(4), 341–359.

Mustak, M., Salminen, J., Plé, L., & Wirtz, J. (2021). Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. *Journal of Business Research*, *124*, 389–404. https://doi.org/10.1016/j.jbusres.2020.10.044

Patel, S. (2020, July 15). *8 Ways to Identify and Fix Customer Pain Points*. REVE Chat. https://www.revechat.com/blog/customer-pain points/

Patnaik, D., & Becker, R. (1999). Needfinding: The Why and How of Uncovering People's Needs. *Design Management Journal (Former Series)*, *10*(2), 37–43. https://doi.org/10.1111/j.1948-7169.1999.tb00250.x

Pawar, U. (2020, December 4). *Lets Open the Black Box of Random Forests*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/12/lets-open-the-black-box-of-random-forests/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Plangger, K., & Montecchi, M. (2020). Thinking beyond privacy calculus: Investigating reactions to customer surveillance. *Journal of Interactive Marketing*, *50*, 32–44.

Price, R., & Wrigley, C. (2016). Design and a deep customer insight approach to innovation. *Journal of International Consumer Marketing*, *28*(2), 92–105.

Price, R., Wrigley, C., & Straker, K. (2015). Not just what they want, but why they want it: Traditional market research to deep customer insights. *Qualitative Market Research: An International Journal*.

Rambocas, M., & Pacheco, B. G. (2018). Online sentiment analysis in marketing research: A review. *Journal of Research in Interactive Marketing*, *12*(2), 146–163. https://doi.org/10.1108/JRIM-05-2017-0030

Rawson, A., Duncan, E., & Jones, C. (2013). The truth about customer experience. *Harvard Business Review*, *91*(9), 90–98.

Ray, S. (2017, September 8). Commonly Used Machine Learning Algorithms | Data Science. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Reinstein, I. (2017). *XGBoost, a Top Machine Learning Method on Kaggle, Explained*. KDnuggets. https://www.kdnuggets.com/xgboost-a-top-machine-learning-method-on-kaggle-explained.html/

Reisenbichler, M., & Reutterer, T. (2019). Topic modeling in marketing: Recent advances and research opportunities. *Journal of Business Economics*, *89*(3), 327–356.

Reisenbichler, M., Reutterer, T., Schweidel, D., & Dan, D. (2021). *Supporting Content Marketing with Natural Language Generation*. 93711.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *3*(22), 41–46.

Rojas, R. (2013). *Neural networks: A systematic introduction*. Springer Science & Business Media.

Rooderkerk, R. P., & Pauwels, K. H. (2016). No comment?! The drivers of reactions to online posts in professional groups. *Journal of Interactive Marketing*, *35*, 1–15.

Rydén, P., Ringberg, T., & Wilke, R. (2015). How managers' shared mental models of business–customer interactions create different sensemaking of social media. *Journal of Interactive Marketing*, *31*, 1–16.

Said, E., Macdonald, E. K., Wilson, H. N., & Marcos, J. (2015). How organisations generate and use customer insight. *Journal of Marketing Management*, *31*(9–10), 1158–1179.

Salian, I. (2018, August 2). *Supervised Vs. Unsupervised Learning*. The Official NVIDIA Blog. https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/

Salminen, J., Almerekhi, H., Dey, P., & Jansen, B. J. (2018, October 15). Inter-rater agreement for social computing studies. *Proceedings of The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018)*. The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018), Valencia, Spain.

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-Centric Computing and Information Sciences*, *10*(1), 1. https://doi.org/10.1186/s13673-019-0205-6

Salminen, J., Kandpal, C., Kamel, A. M., Jung, S., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, *64*, 102771. https://doi.org/10.1016/j.jretconser.2021.102771

Salminen, J., Yoganathan, V., Corporan, J., Jansen, B. J., & Jung, S.-G. (2019). Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type. *Journal of Business Research*, *101*, 203–217.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *ArXiv Preprint ArXiv:1910.01108*.

Sawhney, M., Verona, G., & Prandelli, E. (2005). Collaborating to create: The Internet as a platform for customer engagement in product innovation. *Journal of Interactive Marketing*, *19*(4), 4–17. https://doi.org/10.1002/dir.20046

Schaffhausen, C. R., & Kowalewski, T. M. (2015). Large-scale needfinding: Methods of increasing user-generated needs from large populations. *Journal of Mechanical Design*, *137*(7). https://doi.org/10.1115/1.4030161

scikit-learn. (2020). *f1_score—Scikit-learn 0.24.1 documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

SHAP latest documentation. (2018). Read the Docs. https://shap.readthedocs.io/en/latest/

Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *ArXiv Preprint ArXiv:1803.09820*.

Srivastava, N. (2013). *Improving Neural Networks with Dropout* [Master's thesis]. University of Toronto.

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Straker, K., Mosely, G., & Wrigley, C. (2021). An approach to integrating market research with customer insights through the development of IoT products. *Journal of International Consumer Marketing*, *33*(3), 239–255.

Sugimura, P., & Hartl, F. (2018). Building a reproducible machine learning pipeline. *ArXiv Preprint ArXiv:1810.04570*.

Sun, B., Li, S., & Zhou, C. (2006). "Adaptive" learning and "proactive" customer relationship management. *Journal of Interactive Marketing*, *20*(3), 82–96. https://doi.org/10.1002/dir.20069

Syam, N., & Sharma, A. (2018). Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. *Industrial Marketing Management*, *69*, 135–146. https://doi.org/10.1016/j.indmarman.2017.12.019

Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*, *38*(1), 1–20. https://doi.org/10.1287/mksc.2018.1123

Urban, G. L., & Von Hippel, E. (1988). Lead user analyses for the development of new industrial products. *Management Science*, *34*(5), 569–582.

Verhoef, P. C., Venkatesan, R., McAlister, L., Malthouse, E. C., Krafft, M., & Ganesan, S. (2010). CRM in Data-Rich Multichannel Retailing Environments: A Review and Future Research Directions. *Journal of Interactive Marketing*, *24*(2), 121–137. https://doi.org/10.1016/j.intmar.2010.02.009

Von Hippel, E. (1986). Lead users: A source of novel product concepts. *Management Science*, *32*(7), 791–805.

Von Hippel, E. (2009). Democratizing innovation: The evolving phenomenon of user innovation. *International Journal of Innovation Science*, *1*(1), 29–40.

Wang, B., Miao, Y., Zhao, H., Jin, J., & Chen, Y. (2016). A biclustering-based method for market segmentation using customer pain points. *Engineering Applications of Artificial Intelligence*, *47*, 101–109. https://doi.org/10.1016/j.engappai.2015.06.005

Wang, W., Arya, D., Novielli, N., Cheng, J., & Guo, J. L. C. (2020). *ArguLens: Anatomy of Community Opinions On Usability Issues Using Argumentation Models* (R. Bernhaupt, F. “Floyd” Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjørn, S. Zhao, B. P. Samson, & R. Kocielnik, Eds.; pp. 1–14). ACM. https://doi.org/10.1145/3313831.3376218

Wang, Y., Mo, D. Y., & Tseng, M. M. (2018). Mapping customer needs to design parameters in the front end of product design by applying deep learning. *CIRP Annals*, *67*(1), 145–148. https://doi.org/10.1016/j.cirp.2018.04.018

Weinberg, B. D., de Ruyter, K., Dellarocas, C., Buck, M., & Keeling, D. I. (2013). Destination Social Business: Exploring an Organization’s Journey with Social Media, Collaborative Community and Expressive Individuality. *Journal of Interactive Marketing*, *27*(4), 299–310. https://doi.org/10.1016/j.intmar.2013.09.006

Yang, Y., Zhang, K., & Kannan, P. K. (2021). Identifying Market Structure: A Deep Network Representation Learning of Social Engagement. *Journal of Marketing*, 00222429211033585. https://doi.org/10.1177/00222429211033585

Yiu, T. (2019, August 14). *Understanding Random Forest*. Medium. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Zhang, J., Wei, X., Fukuda, H., Zhang, L., & Ji, X. (2021). A Choice-based conjoint analysis of social media picture posting and souvenir purchasing preference: A case study of social analytics on tourism. *Information Processing & Management*, *58*(6), 102716. https://doi.org/10.1016/j.ipm.2021.102716

Zhang, M., Jansen, B. J., & Chowdhury, A. (2011). Business engagement on Twitter: A path analysis. *Electronic Markets*, *21*(3), 161.

Zhang, Y., Baldridge, J., & He, L. (2019). PAWS: Paraphrase adversaries from word scrambling. *ArXiv Preprint ArXiv:1904.01130*.

Zhou, F., Ayoub, J., Xu, Q., & Jessie Yang, X. (2020). A machine learning approach to customer needs analysis for product ecosystems. *Journal of Mechanical Design*, *142*(1). https://doi.org/10.1115/1.4044435