Text Analysis of Customer Support Tweets

1 INTRODUCTION

In today's highly competitive market landscape, where products and services are becoming increasingly commoditised, customer support often stands out as a crucial tool for companies to distinguish themselves from their competitors. Recent studies reveal that 73% of customers point to customer service as a key factor in their purchasing decisions, with 75% of customers willing to spend more with brands that prioritise customer support over other buying factors, such as price, availability and quality [1, 2].

By resolving issues quickly and efficiently, companies demonstrate a high level of care and commitment to customer satisfaction. This encourages the development of lasting customer-business relationships, ultimately increasing the likelihood of customer loyalty to a brand. Moreover, in today's age of social media and online reviews, positive experiences with customer support leads to invaluable word-of-mouth publicity and enhanced brand reputation. With studies showing that 60% of customers have purchased from a brand solely based on the level of customer support they expect to receive [2]. Therefore, investing in robust customer support isn't just a matter of addressing complaints, it's an investment in the longevity and prosperity of the business itself.

1.1 The Role of Artificial Intelligence

Each year, companies report an increase in the number of customer support inquiries, with 57% of customer support leaders anticipating a further increase of up to 20% in the next 1-2 years [2]. This places a significant burden on customer support staff, resulting in longer wait times for customers and a decline in customer satisfaction. To help manage this increase in demand, many companies are turning to artificial intelligence (AI) to streamline the support processes, with 70% of customer support leaders planning to integrate AI into their support systems within the next two years [2].

In particular, there has been a notable shift towards utilising text analysis to automatically categorise, prioritise and respond to customer inquiries. This shift is driven by evolving consumer preferences, with a growing majority of customers (67%) preferring text-based communication over other traditional methods [3]. Therefore, this project will focus on the development of a text analysis system in line with the changing demands of the customer support industry.

1.2 Similar Research

Text analysis of customer support messages can be broadly split into two main areas: topic modelling and sentiment analysis. Topic modelling focuses on identifying and extracting the main themes or topics present within text, allowing companies to understand the most common issues raised by customers and automatically categorise inquiries for faster resolution by staff [4]. A variety of different tools and methods have been developed to achieve this task, with two of the most common being Top2Vec and BERTopic. These methods both employ similar techniques, utilising sentence embeddings to create vector representations of each document/sentence, which are subsequently clustered and labelled to identify the topics present within a set of documents. However, the key difference between these methods lies in how each cluster is labelled. Top2Vec assigns labels based on the word closest to the cluster's centroid, whereas BERTopic uses a class-based variation of TF-IDF. Both methods have seen high levels of success, with papers by D.Hendry et Al. 2021 and A.Patel et Al. 2023 displaying the creation of high-quality clusters that accurately extract topics within customer support text [5, 6].

On the other hand, sentiment analysis evaluates the emotional tone expressed within text, enabling companies to gauge customer satisfaction and prioritise responses based on the sentiment conveyed. One such tool commonly used within customer support analysis is VADER, which utilises a dictionary of words and a set of rules, such as the use of capitalisation and punctuation, to gauge sentiment strength. This model is also pre-trained on social media text, so can effectively deal with the large variations and errors present within customer messages [7]. For these reasons, VADER often performs to a high standard when analysing sentiment within customer support messages, with a study by V. Nurcahyawati and Z. Mustaffa 2023, resulting in higher accuracy scores when compared to manual annotation [8].

1.3 Aims & Objectives

Many existing projects utilise only one of these methods, focusing solely on either discovering underlying topics or analysing customer sentiment. This approach often results in solutions that fail to extract all relevant information. For instance, tweets that praise Apple for the new iPhone's camera and those reporting a new camera glitch should not be treated equally due to their conflicting sentiment. However, they would likely be clustered together using topic modelling alone. Therefore, by combining both topic modelling and sentiment analysis, companies can gain more valuable insights into customer needs, preferences, and overall opinion. This integration allows for more effective support strategies to be developed in the aim of enhancing customer satisfaction. As such, this project will aim to utilise both techniques in a system that can automatically fulfil the following research objectives:

(1) Cluster customer messages into distinct categories
(2) Validate the clusters are well formed (correctly separate messages by topic)
(3) Label each cluster meaningfully
(4) Rank and prioritise customer messages based on sentiment

The methods used to achieve these objectives are be outlined in Section 2 and their results discussed in Section 3.

2 TECHNICAL SOLUTION

2.1 Dataset

The dataset chosen for analysis in this project is the 'Customer Support on Twitter' dataset, publicly available on Kaggle. It comprises over 3 million customer support tweets from over 100 prominent

brands that encompass a wide variety of customer inquiries, complaints, and responses from real-world customer service interactions. Therefore, this dataset provides a comprehensive representation of customer needs and sentiments, allowing for the exploration of a diverse range of topics and sentiments that accurately reflect the needs of real-world customers.

The dataset is stored as a CSV file, with each row containing a customer tweet and all relevant information, most notably anonymised tweet and author IDs, inbound status (a boolean indicating whether the tweet is directed to or written by the company), the tweet's text, and the recipient's ID. This format allows for straightforward processing and analysis of each company's customer inquiries, aiding in the efficient development of the proposed system and its further application to real-world problems.

2.2 Topic Modelling

After evaluating various topic modelling frameworks, BERTopic emerged as the optimal choice due to the high levels of accuracy and efficiency demonstrated in similar studies [5]. Additionally, BERTopic offers greater customisation compared to similar models such as Top2Vec, thanks to a wider range of pre-trained models available [4].

BERTopic consists of a pipeline of four methods executed sequentially, each of which was implemented individually to allow for higher levels of customisation and fine-tuning, contributing to a higher-performing system. The development of each section of the pipeline is outlined in the following subsections.

2.2.1 *Sentence Embeddings.*

Sentence embeddings are fixed-size, high-dimensional vector representations of sentences generated by deep learning models. They are designed to capture the semantic meaning of sentences, in which vectors with similar values represent sentences of similar topics. This project utilises SBERT, a sentence embedding model selected for its ability to analyse entire sentences rather than their individual words by leveraging multiple (siamese) BERT networks. This architecture results in a high-precision system that runs in a fraction of the time required by traditional BERT models [9]. Furthermore, the pre-trained model chosen for study was 'all-MiniLM-L6-v2' as it offers high performance, whilst also being 5 times faster than models of similar accuracy [10].

2.2.2 *Dimensionality Reduction and Clustering.*

Dimensionality reduction, achieved through the use of UMAP, is a crucial step carried about before clustering. It involves converting sentence embedding vectors from a high number of features to low dimensional vectors of a pre-specified length. Therefore, reducing sparsity within the data, decreasing noise and allowing for well-defined clusters to be found with minimal loss of information [11]. The resultant vectors are then clustered using HDBSCAN, a hybrid algorithm that combines hierarchical and density-based clustering methods. This approach aims to identify clusters of varying shapes and densities while mitigating the influence of outliers likely to be found in social media text [12].

However, for these techniques to generate meaningful clusters, it is essential to determine the ideal parameters for both UMAP and HDBSCAN. Given the high computational cost associated with topic modelling and the numerous parameters required by UMAP and HDBSCAN, this project utilised Bayesian optimisation. In an aim to iteratively find better parameter values by creating a probabilistic model mapping parameters to an objective function [13].

Initially, silhouette coefficient was implemented as the objective function to assess the quality of the clusters formed. However, across multiple runs, consistently low scores, indicating poor clustering, were obtained. This yielded limited useful information for the parameter tuning of the system, likely due to the utilisation of different metrics for evaluating clusters. Specifically, while the silhouette coefficient employs distances to assess cluster quality, HDBSCAN relies on density to define clusters [14]. Therefore, the cluster confidence cost, which calculates the proportion of data points for which HDBSCAN assigns a cluster membership probability greater than 80%, was ultimately chosen as the objective function. In which values closer to zero indicate the presence of higher quality clusters.

2.2.3 *Cluster Labelling.*

To extract topic labels, class-based TF-IDF was implemented, which assesses the significance of a word to a cluster by comparing its frequency within that cluster to its frequency across all clusters. However, before computing this, the text is first preprocessed, removing noise and irrelevant information while preserving the original semantic meaning. The preprocessing steps carried out were as follows:

- Convert all words to lowercase
- Remove stopwords (commonly used words with little useful information)
- Remove punctuation
- Lemmatisation (reduce words to their root synonyms)

Following this preprocessing the top 5 words deemed most significant by cTF-IDF are used to label each cluster.

2.3 Sentiment Analysis

Given the widespread use of VADER for sentiment analysis in similar research projects and its pre-training on social media text. VADER stands out as a particularly suitable model for the Twitter-based dataset used within this project. Therefore, VADER was utilised throughout this project to determine the overall sentiment of customers across each company within the dataset. Additionally, it was also used to further analyse the messages within each cluster to distinguish between genuine customer complaints/inquiries and other messages where the company was merely tagged, minimising time wasted by customer support staff.

3 RESULTS

3.1 Clusters Formed

To evaluate the topic modelling methodology outlined in Section 2.2, clusters were generated and evaluated for the three companies with the highest number of inquiries: Apple, Amazon, and Uber. This evaluation consists of two parts. First, confirming the clusters are

well-defined, characterised by points of high cohesion within each cluster, and second ensuring the effective separation of messages according to their underlying topics. To determine if the clusters are well-defined, two variations of the confidence cost were calculated: one including and one excluding the data points identified as noise by HDBSCAN (cluster -1), accounting for the presence of messages that do not fit into any of the discovered topics. The number of clusters and cluster confidence costs are displayed in Table 1 below.

	Apple	Amazon	Uber
No. Clusters	361	288	189
Cost w Noise	0.30	0.28	0.34
Cost w/o Noise	0.006	0.007	0.013

Table 1: Clusters & Costs

These results clearly demonstrate that the topic modelling methodology, combined with Bayesian optimisation, were able to create a large number of clusters for each company, successfully separating tweets based on information extracted from sBERT. Additionally, the clusters generated are of high quality, with around only 1% of clustered tweets having cluster membership probabilities less than 80%. However, the cost including noise is around 30% for all three companies. Since this cost is significantly higher than that without noise, it indicates a large number of tweets that could not be assigned to any cluster. Such a result is somewhat expected due to the unstructured nature of social media text, which frequently includes irrelevant content, typos, and spam messages that don't align with any recurring underlying topics within the dataset. However, not all tweets identified as noise follow this pattern and it may be the case there aren't enough tweets in a topic to form a new distinct cluster. This is a limitation of the topic modelling methodology used within this project, of which a potential resolution is provided in Section 3.3.

While the evaluation of clustering costs suggests that the clustering has been successful in partitioning the dataset, it is insufficient to conclusively assert that the topic modelling has effectively split the dataset by topic. Therefore, further analysis of the contents within each cluster is essential to provide insights into the coherence and relevance of the clustered messages. To achieve this, each company's inquiries are sorted by their cluster label and written to a JSONlines file, allowing for the efficient analysis of the contents of each cluster. Table 2 below shows some example topics of clusters found.

Apple	Amazon	Uber
iCloud backup	password reset	cleaning fee
iMessage errors	firestick remote battery	phone left in car
weather widget	charged twice	driver insurance

Table 2: Example topics found

The analysis of the generated clusters (including the examples above) further confirms the successful extraction and clustering of the underlying topics within each company's inquiries. This results in a system that can automatically categorise and diagnose the type of issue a customer is facing, leading to quicker response times by staff. Furthermore, by making minor adjustments to the parameters found by Bayesian search, the number and size of clusters can be adjusted, ultimately determining how broad or specific the topic clusters generated are. For example, a cluster regarding the location an Amazon driver left a package can be further divided into clusters including: left with neighbour, in wheelie bin, left on porch. Therefore, the topic modelling methods implemented result in a system that is both accurate and customisable to suit its intended purpose.

Finally, generating the top five words for each cluster using TF-IDF resulted in largely logical labels for the clusters. With clusters regarding linking a Prime account to Twitch and Uber's smelling like smoke being labelled as ['twitch', 'prime', 'connect', 'account', 'link'] and ['smell', 'smoke', 'car', 'driver', 'like'] respectively.

3.2 Sentiment Analysis

Although the topic modelling effectively separates customer messages by topic, there is still the issue of discerning genuine inquiries from other types of customer messages within each topic. For example, although the tweets below are correctly identified as being of the same topic, they should be treated differently by staff due to their sentiment:

"Top customer service today from @AmazonHelp helping me sort out my Super Mario Odyssey pre-order..."

"Are you kidding Amazon? It's almost 4pm. When is Mario Odyssey going to f-ing ship..."

However, by utilising VADER (outlined in Section 2.3) these messages result in sentiment positivity scores of 0.87 and -0.35 respectively. Therefore, by utilising both topic modelling and sentiment analysis, the system developed can effectively separate customer messages by both their topics and sentiment.

3.3 Limitations & Future Work

Although the implemented solution performs effectively for the majority of messages, there are some limitations that result in incorrect topic or sentiment labels. Such limitations and potential future work to resolve them are outlined below:

As discussed in Section 3.1, a considerable portion of messages are labelled as noise as they don't align with any of the extracted topics. While this outcome is somewhat expected, it could lead to the exclusion of less common topics, potentially resulting in a loss of information. However, by rerunning the topic modelling solely on the noise, new clusters can be generated that may extract these less common topics originally identified as noise.

Additionally, although VADER produces accurate sentiment scores for most messages, it often struggles with sarcasm. For example, the sentence "Wow @AppleSupport ! Great job with iOS 11.0.3. It refuses to install on my iPhone 7S Plus. Between this and the buggy Podcasts app, I'm super happy" produces a score of 0.95 as it fails to understand the true intent of the message due to the sarcastic tone conveyed. This is a very common issue among sentiment analysis tools and as such there is currently a large volume of research being undertaken in an effort to create a reasonable solution. Currently, papers such as [15] and [16] have shown improved accuracy in detecting sarcasm through use use of transformer models that could be considered for future application within this project.

REFERENCES

[1] T. Puthiyamadam and J. Reyes, "Future of Consumer Experience Survey 2017/18." PWC, 2018.

[2] Zendesk, "CX Report 2024," 2024.

[3] Kayako, "Live Chat Statistics," https://kayako.com/live-chat-software/statistics/, 2022, [Online; accessed 23-March-2024].

[4] L. Huyberechts, "Topic detection for customer support queries," 2023.

[5] D. Hendry, F. Darari, R. Nurfadillah, G. Khanna, M. Sun, P. C. Condylis, and N. Taufik, "Topic modeling for customer service chats," in *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, 2021, pp. 1–6.

[6] A. Patel, P. Oza, and S. Agrawal, "Sentiment analysis of customer feedback and reviews for airline services using language representation model," *Procedia Computer Science*, vol. 218, pp. 2459–2467, 2023.

[7] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, no. 1, 2014, pp. 216–225.

[8] V. Nurcahyawati and Z. Mustaffa, "Vader lexicon and support vector machine algorithm to detect customer sentiment orientation." *Journal of Information Systems Engineering & Business Intelligence*, vol. 9, no. 1, 2023.

[9] L. Ajallouda, K. Najmani, A. Zellou *et al.*, "Doc2vec, sbert, infersent, and use which embedding technique for noun phrases?" in *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*. IEEE, 2022, pp. 1–5.

[10] M. Nimisha, T. Shamitha, and G. Mishra, "Comparative analysis of embedding models for keyphrase extraction: A keybert-based approach," in *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*. IEEE, 2023, pp. 1–6.

[11] M. Allaoui, M. L. Kherfi, and A. Cheriet, "Considerably improving clustering algorithms using umap dimensionality reduction technique: a comparative study," in *International conference on image and signal processing*. Springer, 2020, pp. 317–325.

[12] L. Wang, P. Chen, L. Chen, and J. Mou, "Ship ais trajectory clustering: An hdbscan-based approach," *Journal of Marine Science and Engineering*, vol. 9, no. 6, p. 566, 2021.

[13] T. T. Joy, S. Rana, S. Gupta, and S. Venkatesh, "Hyperparameter tuning for big data using bayesian optimisation," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 2574–2579.

[14] M. Steinbach, V. Kumar, and P. Tan, "Cluster analysis: basic concepts and algorithms," *Introduction to data mining, 1st edn. Pearson Addison Wesley*, 2005.

[15] S. R. Gosavi, "Transformer based detection of sarcasm and it's sentiment in textual data," Ph.D. dissertation, Dublin, National College of Ireland, 2022.

[16] A. Avvaru, S. Vobilisetty, and R. Mamidi, "Detecting sarcasm in conversation context using transformer-based models," in *Proceedings of the second workshop on figurative language processing*, 2020, pp. 98–103.