



Article

# Extracting Product Improvement Insights from Social Media Comments Using Machine Learning: A Case Study in the Automotive Industry

Philipp Brunner <sup>†</sup>  and Stefanie Vogl <sup>\*,†</sup>

Faculty of Computer Science and Mathematics, University of Applied Sciences Regensburg (OTH Regensburg), 93053 Regensburg, Germany; poststelle@oth-regensburg.de

\* Correspondence: stefanie.vogl@oth-regensburg.de; Tel.: +49-941-943-1306

<sup>†</sup> These authors contributed equally to this work.

## Abstract

This paper presents a scalable machine learning pipeline for extracting actionable, product-related insights from user-generated social media comments. Leveraging sentence embeddings from SBERT and unsupervised clustering (k-Means and agglomerative), the approach structures informal and noisy comments from Instagram and YouTube into topic groups intended to support thematic analysis. A case study on feedback regarding BMW vehicles, comprising more than 26,000 comments, illustrates how the pipeline can reveal recurring user concerns, such as design critiques, usability issues, and technology-related expectations, even in short and unstructured social media comments. The proposed pipeline operates without labeled data or manual annotation, enabling scalable application and transferability across product categories and industries. By transforming large-scale, unstructured consumer feedback into interpretable themes, the pipeline provides product teams with an efficient and structured basis for data-driven product development and improvement.

**Keywords:** social media mining; sentence embeddings; unsupervised clustering; product feedback analysis; SBERT; natural language processing

## 1. Introduction

Developing innovative and customer-centric products is a critical challenge in today's highly competitive and complex automotive industry [1–3] (p. 14). Companies invest heavily to align products with customer needs and to maintain a competitive advantage [4]. In 2021, external market research firms generated approximately \$119 billion globally, including approximately \$29.7 billion in Europe [5], with even higher total expenditures when internal efforts are considered.

While such investments remain essential, the rapidly growing volume of user-generated content on social media offers substantial yet underutilized potential for capturing customer opinions at scale. Platforms such as Instagram and YouTube produce thousands of comments every minute, forming a rich but highly unstructured source of consumer feedback. These comments may reveal product weaknesses, user needs, and concrete suggestions for improvement, often expressed more spontaneously and diversely than in traditional surveys or review platforms. However, the sheer volume, informality, and linguistic variability of this content make it difficult to extract actionable insights in a systematic manner [6].



Received: 7 January 2026

Revised: 29 January 2026

Accepted: 9 February 2026

Published: 11 February 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Sentiment analysis has become a central tool for monitoring user opinions, typically classifying content into positive, negative, or neutral categories [7,8]. While such approaches are useful for assessing general levels of satisfaction, sentiment scores rarely uncover specific user pain points or concrete ideas for product enhancement. In particular, they fail to capture the semantic content required to identify actionable feedback in short, fragmented, and noisy social media text. Extracting such insights therefore requires methods capable of modeling semantic meaning beyond coarse sentiment labels. Recent advances in Natural Language Processing (NLP), Machine Learning (ML), and Large Language Models (LLMs) have made such analyses increasingly feasible.

Against this background, the aim of this study is to develop an automated pipeline that systematically analyzes social media comments to identify product improvement ideas and concrete change requests. In contrast to sentiment-focused approaches, the proposed pipeline targets feedback that can directly inform design and development decisions. To demonstrate its applicability in a real-world context, the methodology is applied to user comments related to BMW AG vehicles, a domain characterized by complex products and highly diverse user perspectives. Rather than introducing new machine learning algorithms, this study integrates established components into a coherent and scalable analysis workflow. The pipeline combines sentence embeddings generated by Sentence-BERT (SBERT) [9], a modification of Bidirectional Encoder Representations from Transformers (BERT) [10], with unsupervised clustering techniques and topic-based summarization. This modular design enables the transformation of noisy user comments into structured thematic insights with minimal manual effort.

The resulting approach is designed with a focus on scalability and practical deployment, while remaining cost-effective and relatively simple to implement for the analysis of unstructured customer feedback. To address the challenges outlined above, this study makes the following main contributions.

#### *Contributions and Research Questions*

First, it proposes an end-to-end machine learning pipeline with human-guided interpretation for extracting actionable, product-related insights from large volumes of unstructured and noisy social media comments. Second, it illustrates how sentence-level semantic embeddings combined with multi-stage unsupervised clustering can structure short, informal user comments into thematic groups that support interpretability, without the need for labeled data or manual annotation. Third, the applicability and practical relevance of the pipeline are examined through a real-world case study in the automotive domain, characterized by complex products and heterogeneous user feedback. Based on these contributions, the study addresses the following research questions (RQs):

- RQ1: To what extent can SBERT-based sentence embeddings support the identification and structuring of semantically related themes in short and noisy social media comments, compared to traditional keyword-based text representations?
- RQ2: How effectively does a multi-stage unsupervised clustering strategy separate actionable product-related feedback from low-value or irrelevant social media content?
- RQ3: Can the proposed pipeline generate meaningful and practically relevant product improvement insights in a complex real-world setting, such as the automotive industry?

## **2. Related Work**

This section reviews prior work related to analytical methods for text mining and semantic analysis, the use of social media data for customer feedback analysis, and challenges associated with complex product domains such as the automotive industry.

### 2.1. Analytical Methods

Early approaches to text mining for product feedback analysis predominantly relied on traditional representations such as Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF) [11,12]. While computationally efficient and easy to implement, these methods are limited in their ability to capture semantic relationships and contextual meaning, particularly in short and informal text.

More recent studies have adopted large pre-trained language models such as BERT to overcome these limitations by incorporating contextual information into text representations. In scenarios involving short, fragmented, and noisy user-generated content, sentence-level representations have proven especially useful. Sentence-BERT (SBERT) extends the BERT framework by optimizing it for the generation of semantically meaningful sentence embeddings, enabling similarity-based analysis and unsupervised clustering of short text units [9]. These advances have significantly improved the ability to structure and analyze informal user feedback beyond keyword-based techniques.

While supervised approaches have shown strong performance in specific text classification tasks, an unsupervised methodology was deliberately adopted. Supervised learning would require manually labeling thousands of comments for every new product domain, which is prohibitively time-consuming and would hinder scalability [13].

### 2.2. Social Media as a Source of Customer Feedback

A substantial body of prior work on product feedback mining is based on structured review platforms such as Amazon or Yelp, where users provide detailed and explicitly product-focused evaluations [14,15]. While such datasets are comparatively homogeneous and easier to process, they primarily reflect the opinions of existing customers and are restricted to specific sales environments.

In contrast, social media platforms provide more spontaneous, diverse, and heterogeneous forms of feedback, including comments from potential customers and non-buyers. Twitter has been widely used for sentiment analysis and opinion mining in this context [16,17]. Studies leveraging platforms such as Instagram and YouTube, however, remain relatively scarce. Existing work on these platforms often focuses on domain-specific applications, including fashion trend analysis on Instagram [18] or sentiment classification of consumer electronics and automotive products on YouTube [19]. Comprehensive approaches that extract structured product-related insights from these highly unstructured and linguistically diverse data sources are still limited.

### 2.3. Application Domain: Automotive Products

Most studies on automated product feedback analysis address relatively simple product categories, such as smartphones, household electronics, or tourism services [7,15,20]. In these domains, user feedback typically refers to well-defined features and is often expressed in comparatively concise and focused language.

The automotive domain presents substantially greater analytical complexity. User comments may address highly technical components (e.g., engine performance or software systems), subjective experiential aspects (e.g., driving experience or aesthetics), or broader brand-related perceptions, frequently expressed in informal, fragmented, or implicit ways. These characteristics make the automated extraction of product-related insights particularly challenging. Only a limited number of studies have explicitly addressed this complexity. For example, Yin et al. [21] analyzed posts from a Chinese automotive forum using BERT-based models to identify product improvement suggestions, highlighting both the potential and the challenges of applying semantic text analysis techniques in this domain.

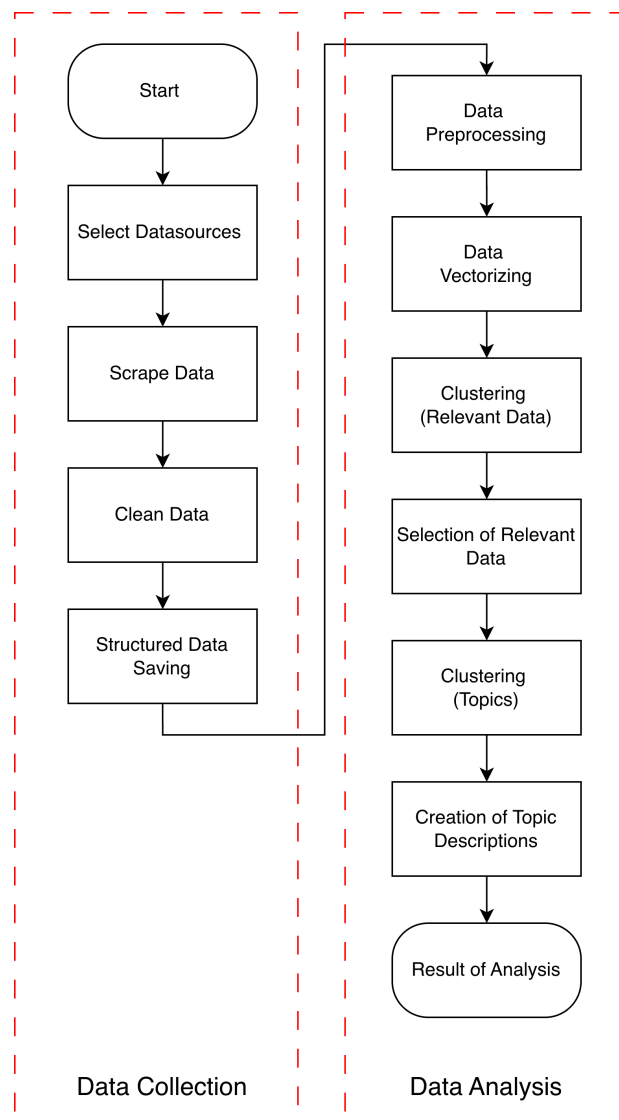
### 3. Materials and Methods

The overall process of data collection and analysis was developed and demonstrated through a case study on BMW vehicle feedback gathered from social media platforms. Based on this case study, the process is designed as a modular, domain-independent, and source-independent analytical workflow that integrates established NLP components in a unified yet adaptable manner. The workflow comprises two main phases: data collection and data analysis.

A detailed overview of all software libraries, tools, and configuration parameters used throughout the pipeline is provided in Table 1.

#### 3.1. Data Collection

The data collection phase involves identifying, acquiring, cleaning, and storing relevant user-generated content from Instagram and YouTube, which as two of the largest social networks produce thousands of comments every minute and together provide both brand-centric and independent user perspectives (see Figure 1) [22]. The choice of data sources strongly influences the diversity and specificity of the collected comments. Certain types of sources, such as review or comparison channels, tend to provide more detailed and feature-focused user feedback.



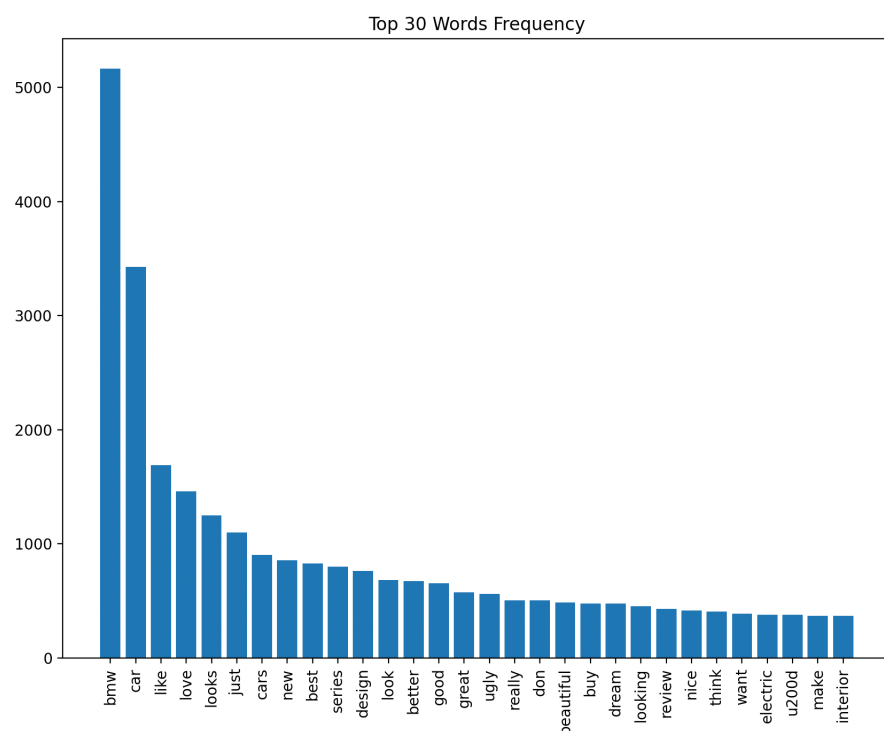
**Figure 1.** Flowchart of data collection and analysis process.

A diverse selection of BMW models was included to capture varied user perspectives across product segments. The dataset comprises the 3 Series, 5 Series, and M3 (representing core and high-performance lines), the i4, i5, and i7 (reflecting electric mobility), and the X5 and iX2 (covering SUV and family segments). For each model, three recent YouTube reviews from creators with over 100,000 followers were selected (24 videos total), along with the ten most recent Instagram posts per model (80 posts total), ensuring a broad and representative dataset.

Data scraping is used to extract the raw content, which is then cleaned to remove inconsistencies and transformed into a structured format suitable for analysis. Cleaning procedures include removing residual HTML elements, parsing metadata, and separating essential variables such as comment text, publication date, and number of likes. Because social media platforms differ significantly in structure and formatting, scraping and cleaning routines must be adapted to each platform. At this stage, preprocessing remains minimal, as the goal is to preserve the raw content while imposing a consistent and traceable data structure. The resulting database contains large volumes of social media comments that form the foundation for subsequent analysis. Prior research indicates that datasets containing between approximately 10,000 and 60,000 short free-text entries are well suited for stable and meaningful clustering when using SBERT-based vectorization [23]. Smaller datasets often lack sufficient semantic diversity and may lead to unstable cluster structures [24].

An analysis of the language distribution in the dataset reveals that approximately 78.7% of all identifiable text content is in English, while the remaining languages are each with individual shares ranging from 0.1% to 1.9%. Given the clear dominance of English and the strong fragmentation of other languages, no language-based filtering or selective translation was applied, and all comments were retained in their original form.

A keyword frequency analysis highlights frequent terms such as “interior,” “design,” and “better,” reflecting common themes of user interest (see Figure 2). These initial observations motivate deeper clustering to extract structured feedback.



**Figure 2.** Top 30 words in dataset.

### 3.2. Preprocessing

The analysis phase begins with preprocessing, where text is refined to ensure high-quality input for downstream tasks (see Figure 1). This includes removing punctuation, emojis, URLs, line breaks, and inline usernames, eliminating empty entries, and applying text normalization (e.g., lowercasing). Emojis are removed to focus the analysis on textual semantic content rather than affective signals. As platform characteristics differ, preprocessing steps must be adapted to each data source.

### 3.3. Vectorization

Earlier research on text mining for product feedback analysis has frequently relied on traditional representations such as Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF) [11,12]. While computationally efficient, these approaches lack semantic and contextual understanding, which limits their ability to capture meaning in short, informal, and noisy social media comments. Similar limitations have been reported for probabilistic topic modeling methods such as Latent Dirichlet Allocation (LDA), which often struggle to identify coherent themes in short texts due to sparse word co-occurrence patterns and limited semantic depth [25].

More recent approaches leverage large pre-trained language models such as BERT to generate contextualized text representations and improve semantic analysis quality [14]. Building on this development, the present study employs Sentence-BERT [9], a BERT-based architecture optimized for producing semantically meaningful sentence-level embeddings. SBERT enables efficient similarity-based analysis and unsupervised clustering of short text units, which is a key requirement for large-scale social media data.

While contemporary Large Language Models (LLMs) provide strong generative and analytical capabilities, their application in large-scale embedding pipelines is often associated with higher computational cost and limited reproducibility. SBERT offers a resource-efficient and cost-effective alternative for sentence embedding, achieving competitive performance for semantic similarity tasks with substantially lower computational overhead [26]. In this work, SBERT embeddings form the basis for all subsequent clustering steps in the proposed pipeline.

### 3.4. Clustering Strategy

The clustering strategy follows a two-stage design, in which an initial noise-reduction step is applied before extracting semantically coherent product-related topics.

#### **Stage 1: Relevance Filtering (Meaningful vs. Low-Value Comments)**

The first stage of the pipeline aims to reduce noise by separating meaningful, product-related feedback from low-value or irrelevant social media content. Low-value comments are understood as entries that do not convey interpretable information about the product itself, such as comments consisting primarily of generic reactions, greetings, or platform-specific expressions.

To this end, sentence embeddings are computed for all comments using SBERT, followed by unsupervised clustering (K-Means and agglomerative). K-Means was selected as a benchmark due to its widespread use and linear-time scalability on large datasets, making it well suited for coarse partitioning. Agglomerative Hierarchical Clustering (AHC), in contrast, does not assume spherical clusters and can reveal nested structures, often resulting in higher semantic coherence in text data [27].

Cluster relevance is determined through qualitative inspection of each cluster's most representative keywords and exemplar comments. Clusters dominated by non-informative patterns (e.g., generic praise or criticism without product reference, or platform-specific noise) are classified as low-value and excluded from further analysis. In contrast, clusters

containing descriptive, evaluative, or comparative statements related to concrete product features, usage experiences, or design aspects are retained.

While this decision process involves human judgment, the same qualitative heuristics are applied consistently across clusters, focusing on semantic informativeness rather than sentiment polarity. This stage substantially reduces noise in the dataset and focuses subsequent analysis on comments that are more likely to contain actionable product improvement insights.

### Stage 2: Topic Clustering

In the second stage, the subset of comments identified as meaningful is further analyzed to extract coherent thematic groups representing recurring product-related topics. Based on the sentence embeddings obtained in the previous step, unsupervised clustering is applied to group semantically similar comments. Both k-Means clustering and agglomerative hierarchical clustering are used to analyze the latent structure of the data and to derive topic-level groupings. The number of clusters is selected empirically using standard heuristics, such as the elbow method [28], in combination with qualitative inspection of cluster coherence. For each resulting cluster, representative comments and frequently occurring terms are examined to support interpretation and to summarize the underlying product-related themes. Cluster labels were generated automatically using KeyBERT, which extracts representative keyword phrases summarizing each cluster's content. These labels provide concise descriptions of the main themes, enabling actionable interpretation of the identified topics.

### 3.5. Reproducibility Details

To ensure transparency and reproducibility of the proposed pipeline, the following implementation details are documented. All parameters listed below correspond to the experimental setup used in the case study.

**Table 1.** Configuration parameters and technical specifications of the proposed analytical pipeline.

| Parameter                                     | Specification Value                                                           |
|-----------------------------------------------|-------------------------------------------------------------------------------|
| Sentence embedding model (SBERT)              | all-distilroberta-v1 <sup>1</sup>                                             |
| Embedding dimension                           | 768                                                                           |
| Distance metric (agglomerative clustering)    | Euclidean                                                                     |
| Linkage (agglomerative clustering)            | Ward                                                                          |
| K-Means parameters                            | init = k-means++, n_init = 10, max_iter = 300, random_state = 42 <sup>2</sup> |
| Vector normalization                          | No                                                                            |
| Dimensionality reduction (visualization only) | t-SNE <sup>3</sup>                                                            |
| Stage 1 clustering method                     | K-Means/agglomerative <sup>4</sup>                                            |
| Stage 2 clustering method                     | K-Means/agglomerative <sup>4</sup>                                            |
| Cluster number selection                      | Elbow method and qualitative inspection                                       |

<sup>1</sup> Pretrained on large-scale social media and web text. <sup>2</sup> Implemented using scikit-learn (version 1.2.2). <sup>3</sup> Applied exclusively for post-hoc visualization of cluster structure. <sup>4</sup> Distinct alternatives evaluated in the analysis.

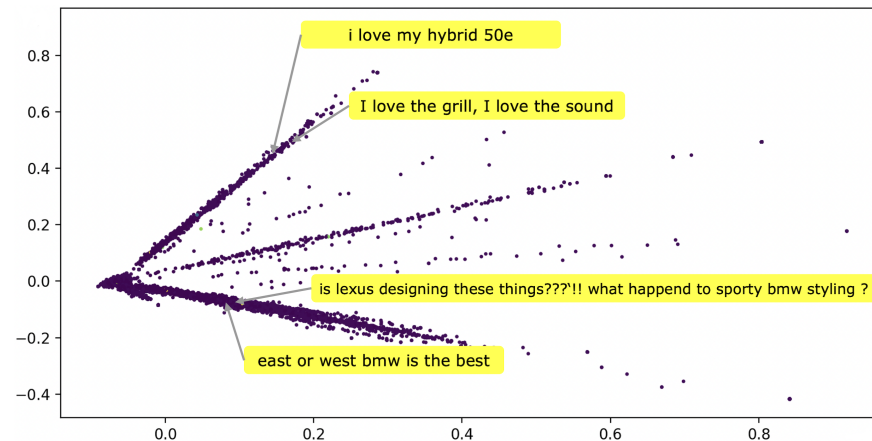
These details enable full replication of the experimental setup.

### 3.6. Ethics and Data Availability

Ethical considerations were addressed by collecting only publicly accessible comments from social media platforms and excluding all user-identifying information such as usernames. No interaction with users was conducted, and no private or restricted data were accessed. According to institutional and national guidelines, the use of publicly available, anonymized data does not require formal ethics approval. The collected dataset is available from the authors upon reasonable request, subject to platform-specific usage policies.

#### 4. Results

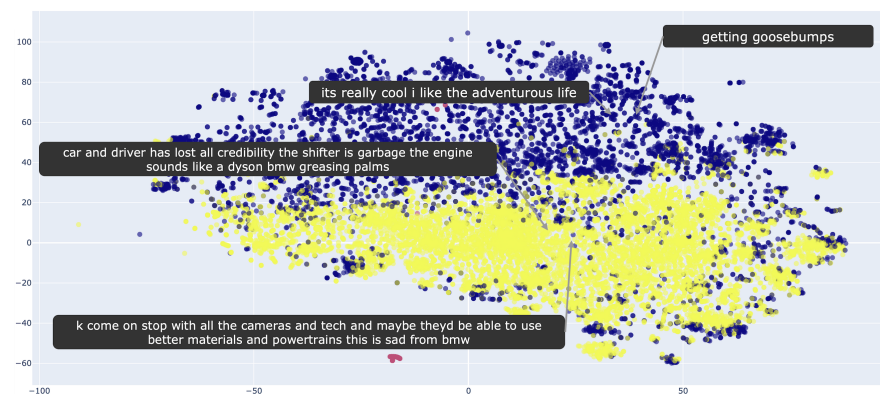
The results focus on the data analysis phase, as it has the strongest influence on the extraction of meaningful insights. Two vectorization methods were tested: a Bag-of-Words (BoW) baseline and SBERT embeddings. The BoW representation showed clear limitations. As illustrated in Figure 3, the resulting vectors were poorly distributed, with many comments collapsing onto a few dominant axes.



**Figure 3.** Data distribution with Bag-of-Words vectorization. Point color holds no meaning. Selected comments are annotated for illustration.

Semantically similar comments were often located far apart, highlighting that BoW relies on lexical overlap and cannot capture contextual meaning, particularly problematic for short, informal social media comments.

In contrast, SBERT produced a well-structured embedding space (Figure 4). Semantically related comments formed coherent neighborhoods, and qualitative inspection of representative comment neighborhoods confirmed that SBERT preserved meaning even when surface wording differed. This richer semantic representation provided a substantially stronger foundation for clustering and was therefore selected for all subsequent steps.



**Figure 4.** Data distribution with SBERT vectorization. Colors indicate cluster assignments from Stage 1 relevance filtering, yellow points represent meaningful comments, blue points represent low-value comments. Selected comments are annotated for illustration.

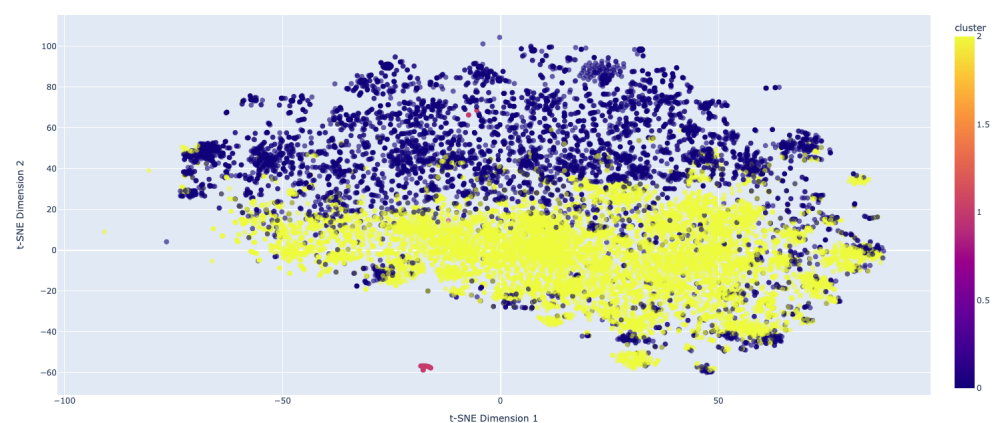
Clustering was performed in two stages. In the first stage, K-Means and Agglomerative Clustering were applied to separate meaningful from low-value comments. The number of clusters was determined using the elbow method. This relevance-filtering step yielded Silhouette Scores of approximately 0.30, indicating moderate within-group similarity and sufficient structure for separating noise from substantive content.

In the second stage, the filtered subset of meaningful comments was clustered into thematic groups, again using both K-Means and Agglomerative variants, with the number of clusters selected via the elbow method. Silhouette Scores in this stage were around 0.04 and changed only marginally across reasonable choices of  $k$ . Such low scores are commonly observed in clustering tasks involving short social media comments embedded in high-dimensional semantic spaces, where substantial topical overlap and shared vocabulary naturally blur cluster boundaries. Despite this, visual inspection of representative clusters and qualitative reading confirmed that cluster cores were thematically coherent, with expected noise at the margins.

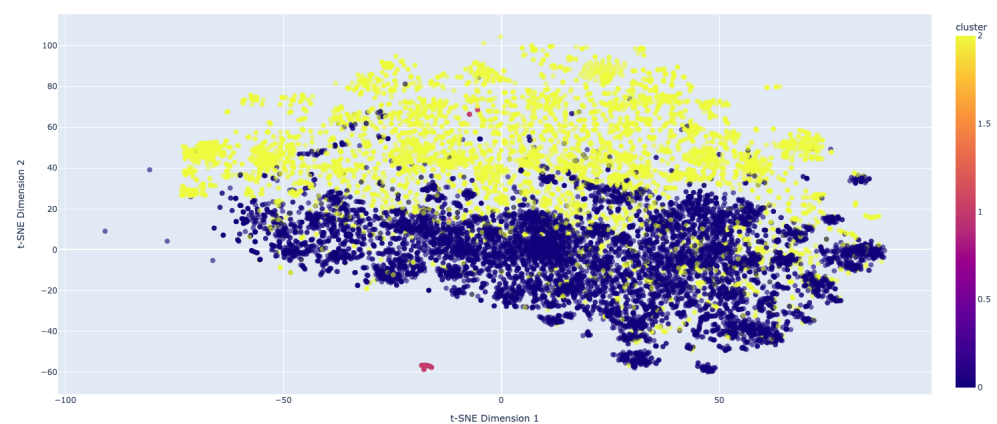
For the intended application of organizing large volumes of user comments into theme-specific pools for downstream expert review, this level of overlap is appropriate, as the overarching thematic structures remain intact.

No separate evaluation of Instagram versus YouTube comments was conducted. Instead, both sources were treated as a unified dataset for all subsequent analyses, as the combined corpus provided a broader and more heterogeneous basis for exploratory clustering.

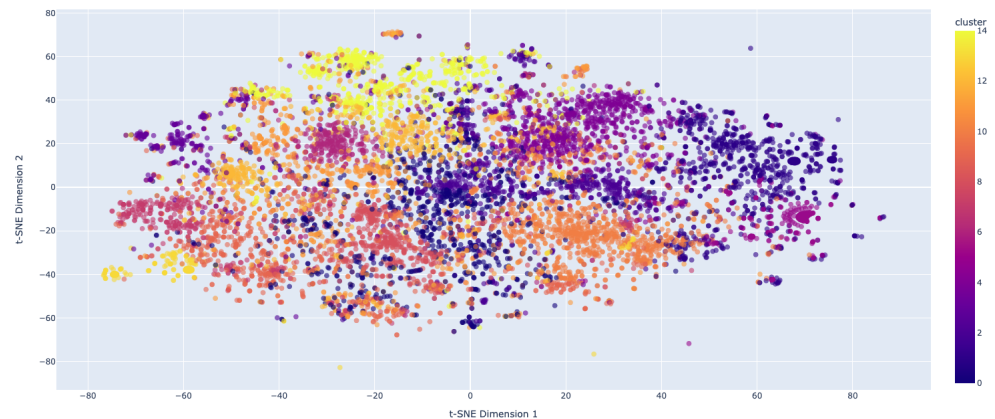
After filtering out irrelevant comments, a comparison of K-Means (Figure 5) and Agglomerative clustering (Figure 6) revealed no substantial differences in the resulting cluster structures. The remaining content-rich subset was then grouped into distinct topic clusters, revealing clear thematic structures within the data. The resulting clusters are shown in Figure 7 and provide an overview of the main discussion topics that emerged from the refined analysis.



**Figure 5.** K-Means clustering on SBERT.



**Figure 6.** Agglomerative clustering on SBERT.



**Figure 7.** Topic clustering with K-Means.

To inspect the cluster structure visually, t-SNE was applied to the SBERT embeddings to project them into two and three dimensions. In the 2D projection, some overlaps between clusters can be observed, whereas the 3D projection reveals clearer separations with several dense cluster cores and partially mixed boundary regions. This pattern is commonly observed in analyses of short social media comments, where high topical similarity naturally leads to soft cluster boundaries and supports the qualitative assessment of clustering structure.

A closer examination of individual clusters indicated strong thematic consistency. For example, one cluster centered on the BMW “kidney” grille feature and captured a wide range of design-related opinions. While some generic statements were present (e.g., “I love the new grille” or “Grille looks terrible”), the dominant portion of the cluster reflected more detailed feedback on proportions, aesthetics, and perceived changes in design direction.

Similarly, clusters related to electric vehicles and overall exterior design showed coherent internal themes. Comments in the EV cluster frequently expressed concerns about weight, pricing, battery resources, or comparisons with combustion engines and competing models. The design-oriented cluster grouped comments highlighting styling preferences, perceived visual appeal, and emotional responses to the aesthetics of specific models.

These excerpts reflect only a small portion of the full dataset, which includes comments across all BMW models selected and both data sources. Accordingly, the linguistic style and comment depth vary considerably across clusters.

The topic descriptions generated in the final step provide concise summaries of each cluster’s dominant themes. For instance, the automatically extracted keywords for the grille cluster (“grill look better”, “larger grills look”, “new grill design”) offer an immediate indication of its content. Across the case study, multiple meaningful clusters were identified, each representing a distinct dimension of user feedback. These structured groupings form an actionable basis for organizing consumer opinions and supporting downstream product improvement analysis.

Taken together, these results illustrate how the proposed pipeline structures large volumes of unstructured social media comments into interpretable thematic groups that can be examined in a downstream analysis.

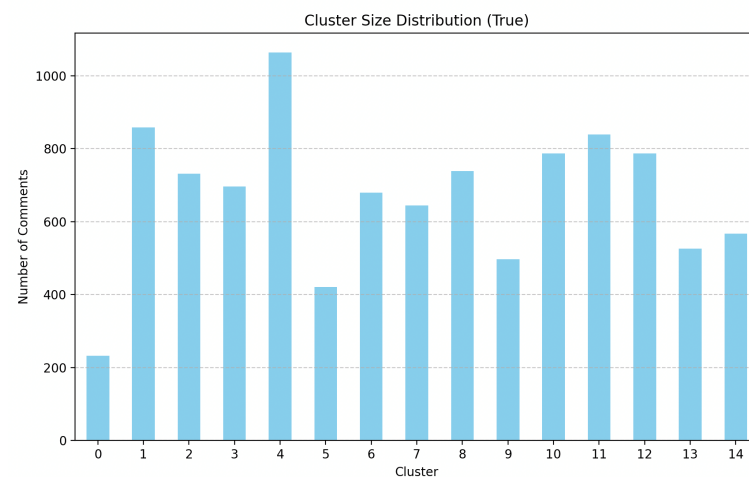
## 5. Discussion

The results demonstrate both the strengths and the limitations of the proposed pipeline for extracting actionable product-related insights from unstructured social media comments. Overall, the approach proves effective in structuring large volumes of noisy, heterogeneous

user-generated content into interpretable thematic clusters, while maintaining scalability and practical applicability.

A central strength of the pipeline lies in its ability to operate without structured input data, predefined categories, or platform-specific metadata. By combining SBERT-based sentence embeddings with unsupervised clustering, the approach consistently yields semantically coherent topic groupings, even in the complex and multi-faceted automotive domain. This indicates that meaningful product feedback can be extracted from short, informal social media comments, supporting the practical relevance of the pipeline for data-driven product development workflows.

The resulting topic clusters exhibit a relatively balanced size distribution, with most clusters containing approximately 600 to 800 comments. Smaller variations can be attributed to natural differences in user engagement across topics rather than to methodological artifacts. As illustrated in Figure 8, no single theme dominates the analysis, suggesting that the pipeline does not disproportionately amplify highly frequent or lexically salient topics at the expense of less prominent but potentially relevant issues.



**Figure 8.** Distribution of comment counts across topic clusters after the second clustering stage.

Despite these strengths, certain steps of the pipeline still require human judgment. In particular, the selection of the number of clusters and the interpretation of the resulting themes involve qualitative assessment, which may affect reproducibility and limit full automation. While qualitative inspection confirmed the interpretability and relevance of the derived topics in this study, future extensions could further reduce manual effort through automated cluster validation measures, assisted topic labeling, or human-in-the-loop evaluation protocols.

Overall, the findings indicate that the proposed pipeline offers a robust and practically viable solution for organizing large-scale social media feedback into actionable thematic structures. The work demonstrates how established methods can be effectively integrated into a coherent end-to-end workflow tailored to the challenges of real-world, noisy social media data.

## 6. Conclusions

This study presented an end-to-end analytical pipeline for extracting actionable product-related insights from large volumes of unstructured social media comments. By combining SBERT-based sentence embeddings with multi-stage unsupervised clustering, the approach demonstrates how noisy and informal user-generated content can be structured into interpretable thematic groups without relying on labeled data or platform-specific metadata.

From a methodological perspective, the contribution of this work lies not in proposing new machine learning algorithms, but in the systematic integration and empirical validation of established components within a coherent and scalable workflow tailored to real-world social media data. The case study in the automotive domain shows that semantic embeddings enable substantially more meaningful structuring of short comments than traditional keyword-based representations, and that a two-stage clustering strategy can effectively separate low-value content from actionable feedback.

From a practical standpoint, the proposed pipeline offers a lightweight and reusable solution for organizations seeking to leverage social media feedback for product development. The reliance on unsupervised methods reduces annotation effort and maintenance costs, while the modular design supports adaptation to different products, platforms, and industries. The results indicate that even highly heterogeneous comment streams can be organized into thematically coherent groups that support expert-driven interpretation and decision-making.

Several limitations of the current study should be acknowledged. First, the identification of meaningful versus low-value content and the selection of the number of clusters involve qualitative judgment, which may affect reproducibility across different analysts or domains. Second, the evaluation of clustering quality is primarily qualitative, reflecting the challenges of applying standard quantitative metrics to short and semantically overlapping social media texts. Third, the empirical validation is limited to a single brand and product category, which constrains direct claims about generalizability.

Future work may address these limitations by incorporating complementary evaluation strategies, such as structured expert validation or alternative cluster quality measures, and by extending the pipeline to additional domains and data sources. Further enhancements could include automated support for cluster selection, multilingual normalization, integration of sentiment or issue prioritization signals, and deployment within an interactive analysis tool for continuous monitoring.

Overall, this work demonstrates that actionable product improvement insights can be systematically extracted from noisy social media data using a carefully designed combination of semantic embeddings and unsupervised learning. The proposed pipeline provides a practical foundation for bridging large-scale consumer feedback and data-driven product development in complex real-world settings.

The results show that SBERT-based sentence embeddings substantially improve the semantic structuring of short and noisy social media comments compared to traditional keyword-based representations (RQ1). The proposed two-stage clustering strategy proved effective in separating low-value content from interpretable, product-related feedback, thereby increasing the usefulness of downstream topic clusters (RQ2). Finally, the automotive case study demonstrates that the pipeline can generate meaningful and practically relevant product improvement insights in a complex real-world setting without relying on labeled data or domain-specific features (RQ3).

**Author Contributions:** Conceptualization, P.B. and S.V.; methodology, P.B. and S.V.; software, P.B.; validation, P.B. and S.V.; formal analysis, P.B.; investigation, P.B.; resources, P.B.; data curation, P.B.; writing—original draft preparation, P.B. and S.V.; writing—review and editing, S.V. and P.B.; visualization, P.B.; supervision, S.V.; project administration, S.V. and P.B.; funding acquisition, S.V. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Publicly available data were analyzed in this study. The data were obtained from the social media platforms described in the text, and each individual comment remains publicly accessible on the respective platform. The full dataset is not stored in a public repository due to privacy regulations (GDPR) and the platforms' Terms of Service, which prohibit the redistribution of scraped user-generated content to third parties or public hosting. These policies are designed to protect users' rights, including the 'right to be forgotten' should they delete their original comments. To ensure compliance with legal and ethical standards, the dataset is available from the corresponding author on reasonable request for research purposes.

**Acknowledgments:** During the preparation of this manuscript, the authors used Large Language Models for translating individual passages from German to English. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Mello, S. *Customer-Centric Product Definition: The Key to Great Product Development*; PDC Professional Publishing: Boston, MA, USA, 2003.
2. Khamidova, F.; Abdurashidova, N.; Khojiev, J.; Makhmudov, S.; Mamadiyrov, Z.; Tukhtabaev, J. Analyzing the Auto Industry: Benchmarking for Competitive Market Assessment. In *Proceedings of the 7th International Conference on Future Networks and Distributed Systems (ICFNDS '23)*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 432–437. [CrossRef]
3. Heneric, O.; Licht, G.; Sofka, W. Challenges and Opportunities for the European Automotive Industry. In *Europe's Automotive Industry on the Move: Competitiveness in a Changing World*; Physica-Verlag HD: Heidelberg, Germany, 2005; pp. 191–205. [CrossRef]
4. Yadav, O.P.; Goel, P.S. Customer satisfaction driven quality improvement target planning for product development in automotive industry. *Int. J. Prod. Econ.* **2008**, *113*, 997–1011. [CrossRef]
5. Statista. Revenue of the Market Research Industry in Europe and Worldwide from 2009 to 2021. Global Market Research. 2022. Available online: <https://de.statista.com/statistik/daten/studie/342892/umfrage/umsatz-von-marktforschungsunternehmen-in-europa-und-weltweit/> (accessed on 29 January 2026).
6. Chan, H.K.; Lacka, E.; Yee, R.W.; Lim, M.K. The role of social media data in operations and production management. *Int. J. Prod. Res.* **2015**, *55*, 5027–5036. [CrossRef]
7. Moro, S.; Rita, P. Data and text mining from online reviews: An automatic literature analysis. *WIREs Data Min. Knowl. Discov.* **2022**, *12*, e1448. [CrossRef]
8. Yue, L.; Chen, W.; Li, X.; Zuo, W.; Yin, M. A survey of sentiment analysis in social media. *Knowl. Inf. Syst.* **2019**, *60*, 617–663. [CrossRef]
9. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 3982–3992. [CrossRef]
10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186. [CrossRef]
11. Abrahams, A.S.; Jiao, J.; Wang, G.A.; Fan, W. Vehicle defect discovery from social media. *Decis. Support Syst.* **2012**, *54*, 87–97. [CrossRef]
12. Su, C.J.; Chen, Y.A. Social Media Analytics Based Product Improvement Framework. In *2016 International Symposium on Computer, Consumer and Control (IS3C)*; IEEE: Xi'an, China, 2016; pp. 393–396. [CrossRef]
13. Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P.; He, L. A Text Classification Survey: From Shallow to Deep Learning. *arXiv* **2020**, arXiv:2008.00364. [CrossRef]
14. Zhai, M.; Wang, X.; Zhao, X. The importance of online customer reviews characteristics on remanufactured product sales: Evidence from the mobile phone market on Amazon.com. *J. Retail. Consum. Serv.* **2024**, *77*, 103677. [CrossRef]
15. Qi, J.; Zhang, Z.; Jeon, S.; Zhou, Y. Mining customer requirements from online reviews: A product improvement perspective. *Inf. Manag.* **2016**, *53*, 951–963. [CrossRef]
16. Lim, K.W.; Buntine, W. Twitter Opinion Topic Model: Extracting Product Opinions from Tweets by Leveraging Hashtags and Sentiment Lexicon. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*; Association for Computing Machinery: Shanghai, China, 2014; pp. 1319–1328. [CrossRef]

17. Hu, G.; Bhargava, P.; Fuhrmann, S.; Ellinger, S.; Spasojevic, N. Analyzing Users' Sentiment Towards Popular Consumer Industries and Brands on Twitter. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*; IEEE: New Orleans, LA, USA, 2017; pp. 381–388. [\[CrossRef\]](#)
18. Hammar, K.; Jaradat, S.; Dokoohaki, N.; Matskin, M. Deep Text Mining of Instagram Data without Strong Supervision. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*; IEEE: Santiago, Chile, 2018; pp. 158–165. [\[CrossRef\]](#)
19. Severyn, A.; Moschitti, A.; Uryupina, O.; Plank, B.; Filippova, K. Multi-lingual opinion mining on YouTube. *Inf. Process. Manag.* **2016**, *52*, 46–60. [\[CrossRef\]](#)
20. Yakubu, H.; Kwong, C.K. Forecasting the importance of product attributes using online customer reviews and Google Trends. *Technol. Forecast. Soc. Change* **2021**, *171*, 120983. [\[CrossRef\]](#)
21. Yin, C.; Jiang, C.; Jain, H.K.; Liu, Y.; Chen, B. Capturing product/service improvement ideas from social media based on lead user theory. *J. Prod. Innov. Manag.* **2023**, *40*, 630–656. [\[CrossRef\]](#)
22. Statista. Biggest Social Media Platforms by Users 2025. We Are Social, DataReportal, Meltwater. 2025. Available online: <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/> (accessed on 29 January 2026).
23. Groot, M.d.; Aliannejadi, M.; Haas, M.R. Experiments on Generalizability of BERTopic on Multi-Domain Short Text. *arXiv* **2022**, arXiv:2212.08459. [\[CrossRef\]](#)
24. Liu, T.; Yu, H.; Blair, R.H. Stability estimation for unsupervised clustering: A review. *Wiley Interdiscip. Reviews. Comput. Stat.* **2022**, *14*, e1575. [\[CrossRef\]](#)
25. Egger, R.; Yu, J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Front. Sociol.* **2022**, *7*, 886498. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Mahajan, Y.; Freestone, M.; Bansal, N.; Aakur, S.; Santu, S.K.K. Revisiting Word Embeddings in the LLM Era. *arXiv* **2024**, arXiv:2402.11094. [\[CrossRef\]](#)
27. Steinbach, M.; Karypis, G.; Kumar, V. A Comparison of Document Clustering Techniques; Technical Report 00-034; University of Minnesota: Minneapolis, MN, USA, 2000. Available online: <https://hdl.handle.net/11299/215421> (accessed on 29 January 2026).
28. Cui, M. Introduction to the K-Means Clustering Algorithm Based on the Elbow Method. *Geosci. Remote Sens.* **2020**, *3*, 9–16.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.