Proceedings of the ASME 2023 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2023
August 20-23, 2023, Boston, USA

IDETC/CIE 2023-115114

EVOLUTIONARY CO-MENTION NETWORK ANALYSIS VIA SOCIAL MEDIA MINING

Phillip A. O. Gavino
Walker Dept. of Mechanical Engineering
The University of Texas at Austin
Austin, TX 78712-1591

Yinshuang Xiao
Walker Dept. of Mechanical Engineering
The University of Texas at Austin
Austin, TX 78712-1591

Yaxin Cui
Dept. of Mechanical Engineering
Northwestern University
Evanston, IL 60208

Wei Chen
Dept. of Mechanical Engineering
Northwestern University
Evanston, IL 60208

Zhenghui Sha*
Walker Dept. of Mechanical Engineering
The University of Texas at Austin
Austin, Texas 78712-1591

ABSTRACT

The immense volume of user-generated content on social media provides a rich data source for big data research. Co-mentioned entities in social media content offer valuable information that can support a broad range of studies, from product market competition to dynamic social network mining and modeling. This paper introduces a new approach that combines named entity recognition (NER) and network modeling to extract and analyze co-mention relationships among entities in the same domain from unstructured social media data. This approach contributes to design for market systems literature because little research has investigated product competition via co-mention networks using large-scale unstructured social media data. In particular, the proposed approach provides designers with a new way to gain insight into market trends and aggregated customer preferences when customer choice data is insufficient. Moreover, our approach can easily support the evolution analysis of co-mention relationships beyond cross-sectional analysis of co-mention networks in a single year due to the abundance of social media data in multiple years. To demonstrate the approach to supporting multi-year product competition analysis, we perform a case study on mining co-mention networks of car models with Twitter data. The result shows that our approach can successfully extract the co-mention relationships of car models in multiple years from 2016 to 2019 from massive Twitter content; and enables us to conduct evolutionary co-mention network analysis with temporal network modeling and descriptive network analysis. The analysis confirmed that the co-mention network is capable of identifying frequently discussed entities and topics, such as car model pairs that often involve in competition and emerging vehicle technologies such as electric vehicles (EV). Furthermore, conducting evolutionary co-mention network analysis provides designers with an efficient way to monitor shifts in customer preferences for car features and to track trends in public discussions such as environmental issues associated with EVs over time. Our approach can be generally applied to other studies on co-mention relationships between entities, such as emerging technologies, cellphones, and political figures.

Keywords: Social media content mining; Named entity recognition; co-mention network; Network evolution.

*Corresponding author: zsha@austin.utexas.edu

1 INTRODUCTION

The growing popularity of social media platforms has led to the accumulation of a vast amount of user-generated content, making it an attractive data source for big data research [1, 2, 3]. The data generated by social media platforms contains a wealth of information, ranging from social connections [4] and public opinions [5] to behavior patterns [6] and engineering design [7, 8, 9], which can be extracted and analyzed for a variety of research purposes. In engineering design, researchers are interested in extracting product feature-related information shared by customers to assist designers in creating products that meet customer preferences. For example, Lim and Tucker developed a Bayesian sampling algorithm to determine optimal search keywords for the accurate extraction of product features from Twitter [8]. In some other domains such as social science, scientists have proposed various approaches to analyze social relationships on social media by modeling followed & following connections or direct communications (e.g., a user posts a direct update to a specific person and engaging in conversation with him/her) [10, 11, 12]. For example, in the study by Huberman et al. (2008) [10], the authors found that the number of friends, had a stronger correlation with user activity than the number of followers. This suggests that when attempting to promote an idea or trend through word of mouth on Twitter, focusing on the number of friends should be the primary strategy.

Besides structured data sources such as the count of followers, post metrics, and comments, there exists additional valuable information embedded within the textual content of social media that remains largely unexplored. One such example is entities co-mentioned in the text, and its theoretical support can trace back to the co-word analysis literature [13, 14]. Researchers revealed that co-word analysis was a powerful tool for discovering associations among research areas in science [13]. In addition, Popovic et al. (2014) extracted the country co-occurrence networks from a large set of financial news and validated its significant overlap with the network built upon the correlation between Credit Default Swaps (CDS) of countries [15]. All of these works demonstrate the value of co-mentioned entity information.

In social media, the entities that are commonly mentioned include people, places, products, organizations/brands, social events, artworks (e.g., music, movies), and terminology [16]. Accordingly, the possible relationships between these entities could be: 1) *association*, denoting that two entities are connected via shared attributes. For example, *My Heart Will Go On* is the theme song for the movie *Titanic*. 2) *Causation*. An example is that the long-term inhalation of specific chemicals is a cause of certain cancers. 3) *Comparison*. For example, two products are co-considered and compared by customers. 4) *Random co-occurrence*, which captures all other undefined relationships. As an example, dealers often announce the arrival of new car models, such as the Toyota Camry, Honda Accord, and Mazda CX-3, and their in-stock status.

Previous research has investigated the potential of co-mention data from different social media platforms, such as car forum posts and Amazon reviews for marketing research [2, 17, 18] and engineering design [19, 20, 21]. The rationale behind these studies is that consumers frequently compare products [22] and their opinions drive content generation and attract visitors [23]. Netzer et al. (2012) work also demonstrated that the co-mention measure from text mining exhibited external validity comparable to that derived from a consumer survey [2]. These findings suggest that co-mention information in user-generated data can provide valuable and reliable insights into product competition for marketing researchers. Meanwhile, knowledge of the structure and evolution of market competition can further benefit the engineering design community [24, 25], as it helps identify customer-desired features through their frequently co-mentioned products.

However, exploring co-mention relationships between entities presents several challenges. First, existing research is based on data collected from specialized platforms, targeting specific groups of users. In contrast, social media platforms (e.g., Twitter) have a more diverse user base and a broader range of product information, but are not fully explored. Second, social media data is often unstructured, making it difficult to accurately extract entity information from short, informal posts that contain emojis, URLs, grammatical errors, and misspellings [26, 27]. The third challenge pertains to characterizing and modeling co-mention relationships and their evolutionary dynamics. Addressing this challenge requires appropriate models to quantitatively and effectively represent the co-mention relationship. To address these challenges, we develop a new approach that integrates named entity recognition (NER) and network modeling to extract and analyze co-mention relationships among entities within the same domain from unstructured social media data. We demonstrate our approach in a case study on co-mention networks of car models from mining multi-year Twitter data.

Compared to existing work, such as the study on co-occurrence networks of car models [2], our study contributes to the literature in two aspects. First, the data source for the existing work is from discussion forums, which are organized by specific topics with more structured and centralized product information. Social media data, on the other hand, could cover discussions beyond pure product comparison, making the extracted co-mention networks more susceptible to noise information. Our approach solved this problem using NER techniques from Natural Language Processing (NLP). Second, this study collects multi-year co-mention network data, enabling the analysis of dynamic competition relations. In particular, we conducted descriptive network analysis on four co-mention networks of car models from 2016 to 2019, thus laying the foundation for our future study on temporal network modeling of product competition in support of design configurations.

The remainder of the paper is organized as follows. In Sec-

tion 2, we introduce our method for identifying product data from social media content using NER and the methods for constructing and analyzing product co-mention networks. Then, we present a case study on car models co-mentioned on Twitter in Section 3. In Section 4, the benefits and limitations of our work as well as its implementation for engineering design are discussed. Finally, in Section 5, we summarize the key findings and suggest future research directions.

2 Research Approach

Figure 1 shows an overview of the proposed approach for co-mention network analysis using social media mining. The approach consists of five steps that are detailed in the following sections.

2.1 Step 1: Social Media Data Collection

Step 1 starts with collecting text data from social media. There are two common tools for data collection: 1) social media application programming interfaces (APIs). Most mainstream social media platforms, including Twitter, Facebook, and Instagram, provide APIs that offer developers and researchers access to various features and data, such as posting updates, accessing user data, reading user profiles, and more. 2) Third-party tools/databases. There are several third-party tools and databases that can help collect data from social media and even provide readily available datasets. One such example is snscrape[1], a Python library that facilitates the scraping of data from multiple platforms.

Due to the massive volume of data on social media, it is impractical to collect all available data. Therefore, it is essential to adopt a data collection strategy, which may vary according to research objectives. For the relationship analysis between co-mentioned entities, one collection strategy is a reference-based keyword search, which involves developing a reference list that encompasses all the named entities that are of interest. The subsequent step entails using these entity names as keywords to filter and extract relevant data from a specific time frame.

2.2 Step 2: Text Data Preprocessing

Step 2 involves pre-processing the obtained textual data, which is crucial for the subsequent analysis. As shown in Figure 1, a typical text data processing pipeline includes six substeps [28, 29], each of which is described below.

1) **Text cleaning**. This involves removing any noisy characters such as punctuation marks, digits, and emojis. This can be achieved by using regular expressions [30].
2) **Tokenization**. This step is to break up the text into tokens where the notion of the token can be individual words, phrases, or other essential units depending on the specific objectives of the task at hand [31].
3) **Stop word removal**. Stop words denote common words that are used for the purpose of joining sentences and do not carry much meaning. Examples include "the," "and," "a," and "an." Removing stop words can reduce the size of the text corpus and herein improve the efficiency of downstream tasks [32].
4) **Stemming / Lemmatization**. Stemming and lemmatization are to normalize words to their base form. Stemming refers to a process that reduces words to their root form by crudely removing ends and derivational affixes. Lemmatization is more advanced and properly reduces words to their base form by removing only inflectional endings based on a vocabulary and morphological analysis [33].
5) **Spell-checking and correction**. Spell-checking and correction is the process of identifying misspelled words in a text and replacing them with their correct spelling [34]. This step is essential for improving the accuracy and readability of the text corpus and the data quality for the downstream tasks.
6) **Removing duplicates**. Removing duplicates involves identifying and removing repeated instances of the same text data within a given dataset or document. This process helps to streamline and simplify text analysis tasks by eliminating redundant information [35].

Although this processing pipeline is general, the inclusion and order of steps may differ depending on different task requirements [29]. For instance, stemming or lemmatization might not be relevant or suitable for performing named entity recognition (NER) on Twitter data, given that such data often comprises informal language and slang that do not conform to conventional grammar rules.

2.3 Step 3: Named Entity Recognition (NER)

In Step 3, the objective is to recognize the named entities of interest that appear together in each text sample, that is, a single post (e.g., one tweet) in the case study. This task, commonly known as Named Entity Recognition (NER) in the field of NLP, can be achieved through two methods. The first method involves importing pre-trained NER models from mainstream NLP libraries such as NLTK [2] and spaCy [3]. However, as these models are typically trained on general data, they may not perform optimally on complex social media data. Therefore, the second method is proposed to train a custom model that can be fine-tuned to better identify entities in specific domains [36].

To create a custom NER model using spaCy, three primary steps must be taken. The first step is to prepare the training and testing data, which includes labeling the training and testing data

[1] Snscrape: https://github.com/JustAnotherArchivist/snscrape

[2] NLTK: https://www.nltk.org/#

[3] spaCy: https://spacy.io/

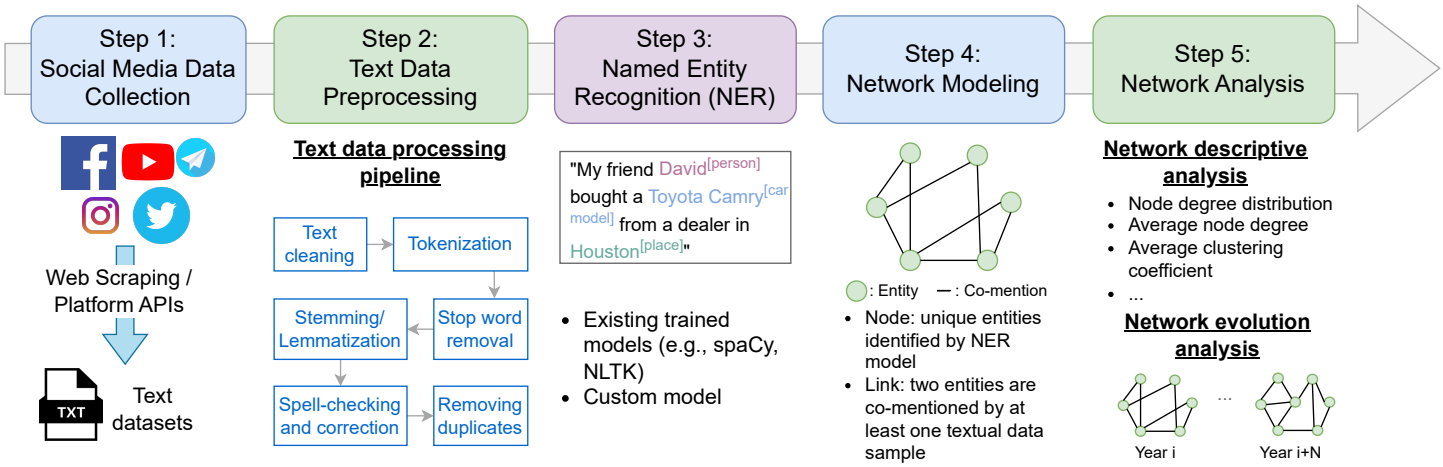

FIGURE 1: Framework of content-based co-mention network mining from social media

and transforming them into the format expected by spaCy. The second step involves modifying the architecture of the spaCy model based on the tasks and specifying the training data and hyperparameters such as batch size and learning rate. In the third step, the model is evaluated using testing data. Evaluation metrics include commonly used machine-learning accuracy scores such as F1-Score, precision, and recall [37].

2.4 Step 4: Network Modeling

In Step 4, the co-mention relationships identified in Step 3 are modeled using a complex network. Complex networks have proven to be a powerful domain-independent representation of complex interactions between entities. For example, social networks have been consistently recognized as an effective means of studying human relationships over the past few decades [38, 39]. Meanwhile, researchers in the engineering domain have also demonstrated the effectiveness of complex networks in system engineering [40] and product design research [41].

Network modeling is a process that encompasses the identification of whether the network is directed or undirected, weighted or unweighted, and the determination of nodes, links, and their corresponding link weights in the case of weighted networks. In the context of co-mention entities, we define nodes as the singular entities identified by the NER model, links as instances where two entities are co-mentioned in at least one text sample, and link weights as the number of times two entities are co-mentioned. Existing work has also used the *lift* value to define link weights [2, 42]. When a causal relationship between entities is unclear, it is reasonable to define the network as undirected.

2.5 Step 5: Network Analysis

In Step 5, following the development of the network model, two types of network analysis are proposed to provide quantitative insight into the co-mention relationships. The first approach involves the application of common network metrics, such as unweighted/weighted degree, network density, global/local average clustering coefficient, and betweenness centrality, to conduct descriptive network analysis [43]. During this process, it is important to connect each metric to the application context for a meaningful interpretation of observed phenomena. For instance, in a co-mention social network, where the co-mentioned entities are individuals' names, individuals with high betweenness centrality can be identified as key connectors between different social clusters.

The second approach is to conduct the network evolution analysis. This is often conducted by performing a time series analysis of the network metrics collected over time and by employing statistical network models to predict network evolution. For example, a temporal exponential random graph model (TERGM) can be used to describe how the probability of edge formation changes over time as a function of various network structures that can incorporate either nodal attributes or edge attributes [24]. A deeper understanding of network evolution obtained from these two types of analysis can support downstream tasks, such as the development of network science-informed deep learning models (e.g., graph neural networks (GNNs)) to predict future co-mention relationships.

3 CASE STUDY

In this section, we present a case study to demonstrate the capability of the proposed approach in co-mention networks of car models using Twitter data in support of the design for vehicle

market systems.

3.1 Step 1: Twitter Data Collection

In this study, we collected data from Twitter based on a reference-based keyword search strategy. To begin, we compiled a list of 949 unique car models from 2010 to 2022 by scraping mainstream model names in English from Cars.com. Then, the third-party tool, snscrape, in conjunction with Twitter internal query search function was utilized to collect tweets from 2016 through 2019. Car models from the reference list were the objects searched for in Twitter's database. To allow for consistent samplings across different time periods, a limit of 20 tweets was collected monthly for each car model. This summed up to 240 tweets per car model, with up to 227,760 in total per year. Due to the lack of tweets on specific car models in some months, the number of tweets was less than the maximum number of tweets that are possibly collected. From 2016 to 2019, the number of tweets collected was 86,962; 90,670; 93,861; and 94,302; respectively. The number of tweets increased over the years, influencing the size of the networks generated during this case study.

3.2 Step 2: Twitter Data Preprocessing

The pipeline used for Twitter text data preprocessing in this study is shown in Figure 2. The first step was removing the URLs from the data frames. This was performed at the outset to simplify the subsequent removal of punctuations. If URLs remained in the data frames, they would be fragmented by the punctuation removal step, rendering their deletion challenging. Then, all punctuation marks were removed. Tokenization was conducted in the third step to split each tweet into individual words. Then, the NLTK library was employed to convert all words to lowercase and remove stopwords, such as "a," "the," and "this". Finally, duplicated tweets were removed from the datasets. These duplicates were deemed to have a high probability of being tweeted by bots whose content was meaningless [44]. After deleting duplicates, the total number of tweets kept was 34,278; 36,940; 43,347; and 49,895 from 2016 to 2019, respectively.

3.3 Step 3: Named Entity Recognition (NER) for Twitter Data

To start this process, we first generated training and testing data using the NER Annotator[4], an online annotation tool, to manually mark the car models in tweets that were processed through Step 2 with the label "CAR." This marking method identifies the beginning and ending indices of each entity in a tweet and subsequently converts this information, along with the tweet content, into the format that is expected by spaCy. We used 2,003

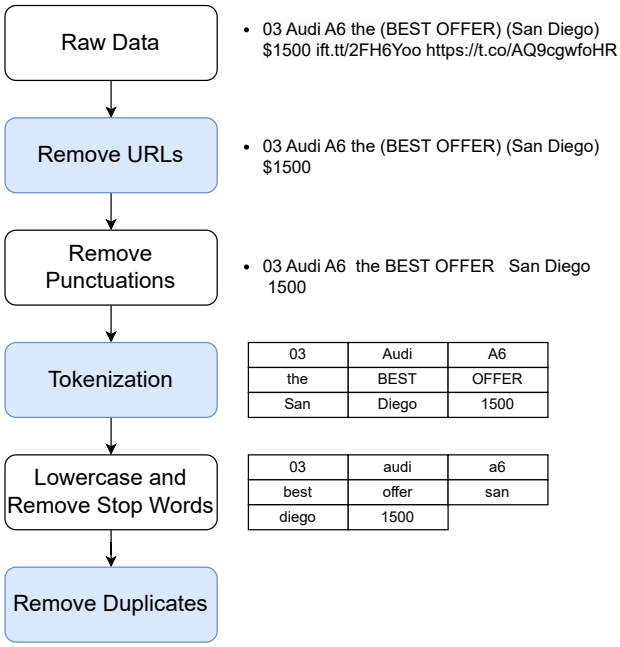

FIGURE 2: Flowchart of the preprocessing method

marked tweets from 2018 as training data and 189, 195, 193, and 193 tweets from 2016 to 2019 as testing data, respectively.

The NER model was then trained using the annotated training data. Upon completion of the training phase, model performance was evaluated using independent test data sets spanning 2016 to 2019. The results of these evaluations were tabulated in Table 1. The results demonstrate that the model achieved F1-scores greater than 69% over the four years, despite being trained solely on data from 2018. Additionally, all precision values were found to be higher than 74%, indicating that more than 74% of the car models recognized by the NER model were correct identification. Furthermore, all recall values exceeded 66%, signifying that the NER model successfully extracted more than 66% of the ground-truth car models (all the manually marked car models within the testing data) from tweets. Finally, we refer to the results of the Twitter Named Entity Recognition shared task associated with the second Workshop on Noisy User-generated Text (W-NUT 2016) to gain general insights into the overall performance of our model. Their results were generated by ten teams. The average F1-Score of these ten NER models was 38.19%, and the highest value was 52.41% [45]. Our model achieved an F1-Score that is approximately 20% higher than the highest F1-Score of the Twitter Named Entity Recognition shared task, justifying the reliability of our trained model.

3.4 Step 4: Twitter Co-Mention Network Modeling

In the process of co-mention network modeling, the cleaned four-year tweet data were processed through the trained NER

[4]NER Annotator: https://tecoholic.github.io/ner-annotator/

TABLE 1: The testing results of NER model by year

Year	F1-Score	Precision	Recall
2016	73.25%	80.42%	67.26%
2017	71.50%	77.67%	66.23%
2018	74.83%	74.03%	75.67%
2019	69.96%	74.37%	66.04%

model to identify car model names in each tweet. Subsequently, only tweets containing more than one car model were retained. The resulting count of retained tweets was 4,747; 6,040; 8,408; 11,220 for the years 2016 to 2019, indicating that the percentages of tweets collected that co-mentioned at least two car models were 13.85%, 16.35%, 19.40%, and 22.49%, which is below 50%. This suggests that the co-mention information of cars is dispersed throughout Twitter. Next, given that there existed multiple variant names for some car models in the extracted model name sets, e.g., Ford F 150 being called "Ford F150", "F 150 Ford", and "FordF150", etc., we only generate co-mention connections between identified car models with names consistent with our reference list from Cars.com, resulting in partial entity information loss. For example, we identified three models from one tweet, including "F150 Ford", "Toyota Highlander", and "Subaru Crosstrek". But given that "F150 Ford" differed from the name "Ford F 150" that was recorded in our reference list, we thereby only generated a co-mention connection between Toyota Highlander and Subaru Crosstrek based on this tweet. [5] Our decision to prioritize an accurate network model over one with more information but greater noise reflects a trade-off. In future work, our aim is to develop a more robust similarity algorithm to address this limitation.

Figure 3 illustrates a co-mention network model based on three annotated tweets. The nodes represent unique car models, links signify two car models being co-mentioned in at least one tweet, and link weights denote the total number of tweets that co-mentioned any two models. Note that no sentiment analysis is conducted in this study; thus,the possible relationship between these comorbid car models could be all four possible relationships stated in Section 1, i.e., association, causation, comparison, and random co-occurrence.

3.5 Step 5: Twitter Co-Mention Network Analysis

Figure 4 visualizes the co-mention networks from 2016 to 2019, and the corresponding unweighted degree distributions are

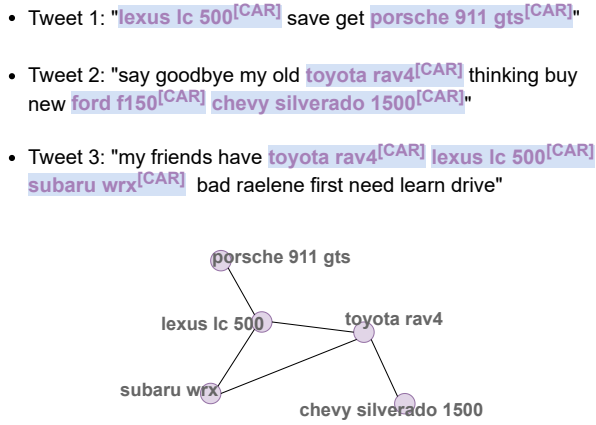

FIGURE 3: An example of co-mention network modeling. These tweets displayed here have undergone processing following Step 2. Nodes are unique car models that were mentioned by all three tweets, and links denote co-mention relationships. For example, a link is built between lexus lc 500 and porsche 911 gts because they were co-mentioned by Tweet 1. We did not include "ford f150" in the network modeling because it is inconsistent with the recalled name listed as "ford f 150" in the reference list.

provided in Figure 5. According to Figure 4, we can observe that the dimensions of the network, including the number of nodes and links, have experienced a marked expansion from 2016 to 2019. This augmentation in the count of nodes signifies an escalation in the number of car models that were co-mentioned in tweets, which can be explained by the number of existing unique car models being cumulative with time. The increasing number of links indicates that more car models are connected through co-mentions. This rise in connections may be the result of a growing interest in car-related topics, such as new automotive technologies (e.g., the surging popularity of electric vehicles), leading to increased discussion.

Furthermore, Figure 5 shows that the degree distributions for both 2016 and 2017 exhibit skewness with a long tail, implying a power law distribution characterized by the fitting curves with R^2 values of 0.862 and 0.812. This suggests the presence of network hubs and the fact that most nodes have relatively few connections. In other words, in 2016 and 2017, a small number of car models were frequently co-mentioned, while the majority of car models (more than 30%) were only co-mentioned once. In contrast, the degree distributions of the 2018 and 2019 networks are spread out (i.e., less skewed than those in 2016 and 2017), indicating that more number of car models were co-mentioned. One possible explanation for this trend is the increase in the overall number of popular car models and heated discussions fueled by the boom of electric vehicles.

By examining the top five hubs, i.e., the top five car models

[5] We utilized text cosine similarity algorithm [46] to detect car models and their variants. However, this algorithm is not robust against some models, such as Nissan Z which was difficult to distinguish from other Nissan models due to its name in the short letter "Z".

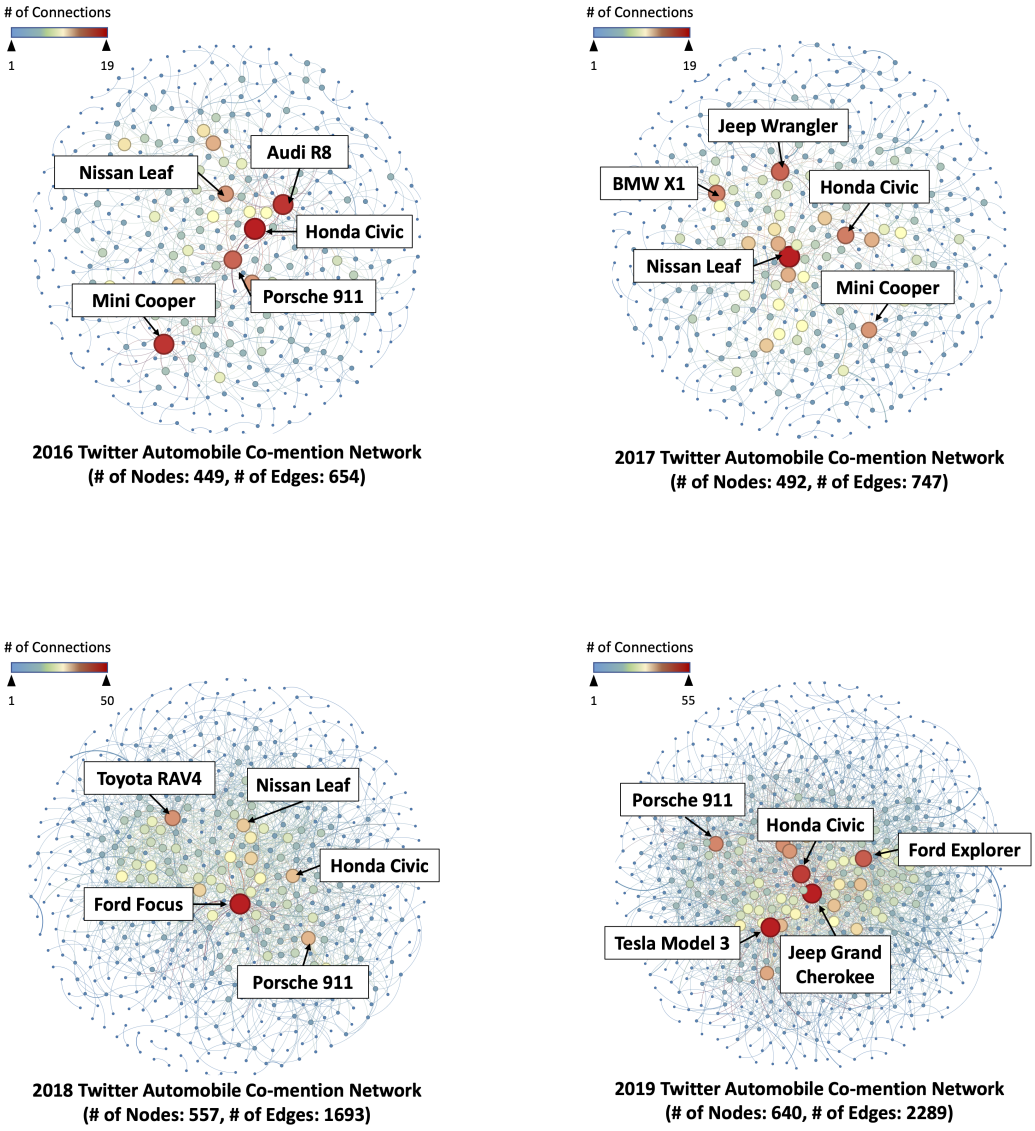

FIGURE 4: Visualizations of co-mention networks from 2016 to 2019. The labeled nodes represent the top-five car models with the most connections in each network, and the corresponding rank information is presented in Table 2.

with the highest unweighted degree, from 2016 and 2019, we observe the evolution of the most frequently discussed car models over time. As presented in Table 2, it is observed that the top five car models vary over time, with Honda Civic being the only model that remains on the list throughout the four years. We also checked the weighted versions of the top five and discovered that the top-one models remain unchanged. For example, Tesla Model 3 holds the top position in both the unweighted and weighted lists in 2019, with an unweighted degree of 55 and a weighted degree of 79. This suggests that Tesla Model 3 was mentioned alongside 55 distinct models and had a total of 79 co-mentions.

Furthermore, we compared the car models that were co-mentioned on social media and their performance in the actual

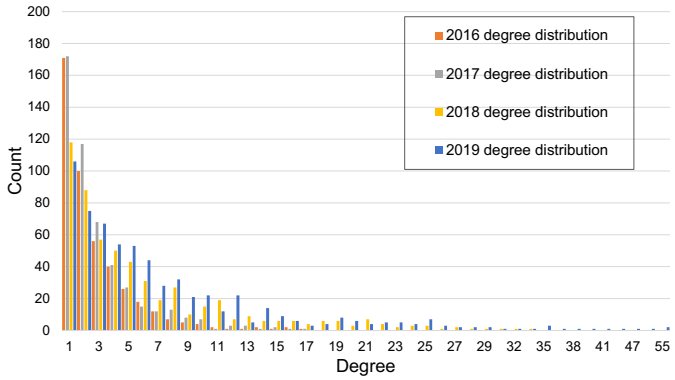

FIGURE 5: Unweighted degree distribution of co-mention networks from 2016 to 2019

sales market, using data of the top five cars with the best sales in 2016 and 2017 from Cars.com [47, 48]. It was observed that the top five best-selling cars were entirely different from the popular cars that were identified in our co-mention networks, indicating whether or not a car model is widely discussed on social media does not necessarily correlate with its sale performance. One possible reason could be that human behavior on social networks can be irrational while purchasing behavior in the real world tends to be more rational and informed by rigorous comparisons. This discrepancy in economics is often described as "talk is cheap." [49]. Another possible reason could be that social media platforms have a diverse user base beyond actual buyers. Hence, while a product or a brand may generate a lot of topics on social media, only a small percentage may come from potential buyers. This further underscores the importance of performing sentiment analysis in our future work, which will allow us to have a better estimate of the type of co-mention relations.

Finally, we identified the car pairs most frequently mentioned in each year, as shown in Table 3, and computed the network metrics associated with each network in Table 4. The results in Table 3 indicate that our approach can successfully capture the car models that were frequently co-mentioned in particular market segments. For example, we found the co-mention of luxury midsize SUVs, BMW X5 vs. Volvo XC90 in 2016, the co-mention of electric vehicles, Tesla Model 3 vs. Chevrolet Bolt EV in 2017, and the co-mention of Jeeps Wrangler and Wrangler Unlimited in 2019. In addition, the trend of the market can also be inferred from the data. For example, the frequent association between Tesla Model 3 and Chevrolet Bolt EV in 2017 serves as evidence of the increasing popularity of electric vehicles.

Regarding the results of the network metrics, the density of a network refers to the proportion of actual connections in a network compared to the total number of possible connections. The reported densities of the co-mention networks are very low (less than 1% on average), indicating sparse car co-mention relations in Twitter data. The weighted and unweighted average degrees measure the average number of co-mentioning occurrences and the average number of co-mentioned car models of a car. The results show that both measures have increased from 2016 to 2019. The local clustering coefficient is computed for each node and indicates how likely the neighbors of a node are also connected. It measures a network's local link density: The more densely interconnected the neighborhood of a node, the higher its local clustering coefficient. The average local clustering coefficients capture the degree of clustering of a whole network by taking the average of local clustering coefficients. The increased local clustering coefficient observed in 2018 and 2019 suggests that there was a greater tendency for a group of car models to be discussed together, compared to 2016 and 2017. This trend may be attributed to the segmentation of car-related discussions on Twitter, where certain car models from the same segment are discussed more frequently. For instance, the co-occurrence of SUV vs. SUV, rather than SUV vs. Sedan, could reflect the clustering of discussions around SUVs during this period.

4 DISCUSSION

In this section, we discuss two benefits of evolutionary co-mention network analysis for product designers and marketing stakeholders, two limitations of this study that motivate our future research, as well as the potential of the proposed approach in support of product design.

Benefits There are two implications of the evolutionary co-mention network analysis. First, the analysis can help designers track changes in customer preferences for car features and emerging technology. In particular, our analysis indicates an increasing co-mention of electric vehicles from 2016 to 2019. This could be due to the rapid development of EV technologies in the last decade which has attracted a lot of attention on social networks. Second, co-mention relations between brands can provide insights into consumer perceptions and potential market competition structures [2]. Although our current results could not directly imply competition relations, the data generated from this preliminary work provide a foundation for our future work on dynamic competition analysis.

Limitations The current study has some limitations that need to be addressed in our future work. As highlighted in Section 3.4, the primary drawback of the proposed approach is its inability to standardize the names of the vehicle models extracted. This has led to the presence of multiple variant names for certain car models, which could introduce noise into our network. To mitigate this issue, we opted to incorporate only models that have names included in our reference list. While this decision allows us to maintain greater consistency and accuracy in our analysis, it sacrifices some co-mention relationships, making them left un-

TABLE 2: Top five car models with the largest unweighted degree (UWD)

2016		2017		2018		2019	
Model	UWD	Model	UWD	Model	UWD	Model	UWD
Honda Civic	19	Nissan Leaf	19	Ford Focus	50	Jeep Grand Cherokee	55
Mini Cooper	18	Jeep Wrangler	16	Toyota RAV4	37	Tesla Model 3	55
Audi R8	18	Honda Civic	15	Porsche 911	33	Honda Civic	51
Porsche 911	16	BMW X1	15	Honda Civic	32	Ford Explorer	47
Nissan Leaf	14	Mini Cooper	14	Nissan Leaf	31	Porsche 911	42

TABLE 3: The most frequently co-mentioned car pairs by year

Year	Linked Car Models	# of Co-mentions
2016	BMW X5 vs. Volvo XC90	6
2017	Tesla Model 3 vs. Chevrolet Bolt EV	8
2018	Buick Envision vs. Cadillac CT6	13
2019	Jeep Wrangler vs. Jeep Wrangler Unlimited	11

TABLE 4: Twitter co-mention network metrics by year

Year	**2016**	**2017**	**2018**	**2019**
Density	0.007	0.006	0.011	0.012
Unweighted Avg. Deg.	2.913	3.037	6.079	7.466
Weighted Avg. Deg.	3.523	3.793	8.047	9.934
Avg. Local Cluster Coeff.	0.125	0.128	0.229	0.249

mined. In the future, a robust text similarity algorithm is required to address this issue. A second limitation of this study is that we did not classify the co-mention relationships into more granular categories or perform spam detection. As a result, our generated networks may contain instances of *random co-occurrence* relationships or advertising information. To achieve a more impartial understanding of the co-mention relationships among car models, additional noise removal techniques may be required to enhance the precision of our findings.

Implementation for engineering design As illustrated in Figure 6, the mined co-mention network from social media data enables us to uncover potential competition relationships between products. By conducting sentiment analysis, we can categorize the co-mention relationships and extract a sub-network dedicated to product competition. In our previous studies, similar co-consideration networks generated by customer choice sets have proven to be effective for demand analysis, which can inform better design decisions [24, 41, 42]. For instance, in a case study on the vehicle market, an exponential random graph model (ERGM) was developed by incorporating the competition network data and the node (car) attributes data such as engine power to estimate the effects of different attributes on the formulation and evolution of competitive relationships. The estimated ERGM can be further integrated into the decision-based design (DBD) framework [50, 51] to use optimization to determine the preferred design alternative. Therefore, the proposed co-mention network mining approach from social media data, i.e., the starting point of this competition network-based design framework, plays a crucial role in providing the competition network input for enterprise-driven design decisions.

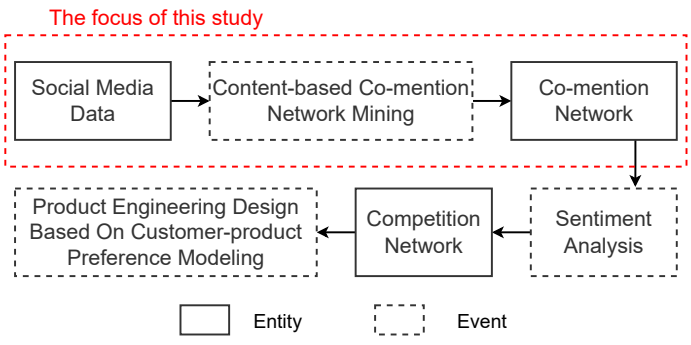

FIGURE 6: The role of this work in competition network-based design framework

5 CONCLUSION

This paper proposes an approach that combines NER and network modeling to extract and analyze co-mention relations among entities embedded in social media data. The proposed method is demonstrated using a case study on a car model co-mention network. Despite the challenges posed by the unstructured and noisy nature of Twitter data, our NER algorithm achieved an F1 score of 70%, a performance that is better than the state-of-the-art. The evolving network models showed that the co-mention network is capable of identifying the competing car models that are frequently discussed on social networks and trending vehicle technologies, such as electric vehicles. Furthermore, the results indicate that the popularity of a product on social media may not necessarily indicate its sales performance. This study offers a unique data mining approach for marketing researchers and product designers in the study of the evolution of marketing structures (competition networks).

In future work, we plan to improve our approach by developing a robust text similarity algorithm that can effectively deal with variant names for specific car models. Also, we plan to carry out a sentiment analysis on social media data to detect spam and classify co-mention relationships into more detailed categories. Lastly, we intend to employ temporal network modeling techniques such as TERGM to identify key factors that affect the formation of co-mention relationships and predict products' future co-mention networks in support of product design and development.

ACKNOWLEDGMENT

The authors acknowledge collaborators Noshir S Contractor, Johan Koskinen, Neelam Modi, and Jonathan Haris Januar for their inputs provided during research meetings. We also greatly acknowledge the funding support from NSF CMMI #2005661 and #2203080.

REFERENCES

[1] Zafarani, R., Abbasi, M. A., and Liu, H., 2014. *Social media mining: an introduction*. Cambridge University Press.

[2] Netzer, O., Feldman, R., Goldenberg, J., and Fresko, M., 2012. "Mine your own business: Market-structure surveillance through text mining". *Marketing Science,* ***31***(3), pp. 521–543.

[3] Aichner, T., Grünfelder, M., Maurer, O., and Jegeni, D., 2021. "Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019". *Cyberpsychology, behavior, and social networking,* ***24***(4), pp. 215–222.

[4] del Fresno García, M., Daly, A. J., and Segado Sanchez-Cabezudo, S., 2016. "Identifying the new influences in the internet era: Social media and social network analysis.". *Revista Española de Investigaciones Sociológicas*(153).

[5] Jungherr, A., 2016. "Twitter use in election campaigns: A systematic literature review". *Journal of information technology & politics,* ***13***(1), pp. 72–91.

[6] De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E., 2021. "Predicting Depression via Social Media". *Proceedings of the International AAAI Conference on Web and Social Media,* ***7***(1), Aug., pp. 128–137.

[7] Stone, T., and Choi, S.-K., 2013. "Extracting consumer preference from user-generated content sources using classification". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 55881, American Society of Mechanical Engineers, p. V03AT03A031.

[8] Lim, S., and Tucker, C. S., 2016. "A bayesian sampling method for product feature extraction from large-scale textual data". *Journal of Mechanical Design,* ***138***(6).

[9] Singh, A. S., and Tucker, C. S., 2015. "Investigating the heterogeneity of product feature preferences mined using online product data streams". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 57083, American Society of Mechanical Engineers, p. V02BT03A020.

[10] Huberman, B. A., Romero, D. M., and Wu, F., 2008. "Social networks that matter: Twitter under the microscope". *arXiv preprint arXiv:0812.1045*.

[11] Tang, L., and Liu, H., 2011. "Leveraging social media networks for classification". *Data Mining and Knowledge Discovery,* ***23***, pp. 447–478.

[12] Yoo, E., Rabinovich, E., and Gu, B., 2020. "The growth of follower networks on social media platforms for humanitarian operations". *Production and Operations Management,* ***29***(12), pp. 2696–2715.

[13] He, Q., 1999. "Knowledge discovery through co-word analysis". *Libr. Trends*.

[14] Chen, X., Chen, J., Wu, D., Xie, Y., and Li, J., 2016. "Mapping the research trends by co-word analysis based on keywords from funded project". *Procedia computer science,* ***91***, pp. 547–555.

[15] Popović, M., Štefančić, H., Sluban, B., Kralj Novak, P., Grčar, M., Mozetič, I., Puliga, M., and Zlatić, V., 2014. "Extraction of temporal networks from term co-occurrences in online textual sources". *PloS one,* ***9***(12), p. e99515.

[16] Room, C., 2020. "Named entity recognition". *Algorithms,* ***8***(3), p. 48.

[17] Won, E. J. S., Oh, Y. K., and Choeh, J. Y., 2018. "Perceptual mapping based on web search queries and consumer forum comments". *International Journal of Market Research,* ***60***(4), pp. 394–407.

[18] Won, E. J., Oh, Y. K., and Choeh, J. Y., 2023. "Analyzing

competitive market structures based on online consumer-generated content and sales data". *Asia Pacific Journal of Marketing and Logistics,* ***35***(2), pp. 307–322.

[19] Jin, J., Ji, P., and Gu, R., 2016. "Identifying comparative customer requirements from product online reviews for competitor analysis". *Engineering Applications of Artificial Intelligence,* ***49***, pp. 61–73.

[20] Suryadi, D., and Kim, H. M., 2019. "A data-driven methodology to construct customer choice sets using online data and customer reviews". *Journal of Mechanical Design,* ***141***(11).

[21] Joung, J., and Kim, H. M., 2021. "Approach for importance–performance analysis of product attributes from online reviews". *Journal of Mechanical Design,* ***143***(8).

[22] Pang, B., Lee, L., et al., 2008. "Opinion mining and sentiment analysis". *Foundations and Trends® in Information Retrieval,* ***2***(1–2), pp. 1–135.

[23] Schindler, R., and Bickart, B., 2005. *Published word of mouth: Referable, consumer-generated information on the internet*. Lawrence Erlbaum Associates, Jan., pp. 32–57.

[24] Xie, J., Bi, Y., Sha, Z., Wang, M., Fu, Y., Contractor, N., Gong, L., and Chen, W., 2020. "Data-driven dynamic network modeling for analyzing the evolution of product competitions". *Journal of Mechanical Design,* ***142***(3), p. 031112.

[25] Wang, M., Sha, Z., Huang, Y., Contractor, N., Fu, Y., and Chen, W., 2018. "Predicting product co-consideration and market competitions for technology-driven product design: a network-based approach". *Design Science,* ***4***, p. e9.

[26] Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., and Lee, B.-S., 2012. "Twiner: named entity recognition in targeted twitter stream". In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 721–730.

[27] Suman, C., Reddy, S. M., Saha, S., and Bhattacharyya, P., 2021. "Why pay more? a simple and efficient named entity recognition system for tweets". *Expert Systems with Applications,* ***167***, p. 114101.

[28] Pradha, S., Halgamuge, M. N., and Vinh, N. T. Q., 2019. "Effective text data preprocessing technique for sentiment analysis in social media data". In 2019 11th international conference on knowledge and systems engineering (KSE), IEEE, pp. 1–8.

[29] Kathuria, A., Gupta, A., and Singla, R., 2021. "A review of tools and techniques for preprocessing of textual data". *Computational Methods and Data Engineering: Proceedings of ICMDE 2020, Volume 1*, pp. 407–422.

[30] Goyvaerts, J., 2017. "Regular expression tutorial-learn how to use regular expressions". *Available online: www. regular-expressions. info (accessed on 31 October 2021)*.

[31] Webster, J. J., and Kit, C., 1992. "Tokenization as the initial phase in nlp". In COLING 1992 volume 4: The 14th international conference on computational linguistics.

[32] Nothman, J., Qin, H., and Yurchak, R., 2018. "Stop word lists in free open-source software packages". In Proceedings of Workshop for NLP Open Source Software (NLP-OSS), pp. 7–12.

[33] Schütze, H., Manning, C. D., and Raghavan, P., 2008. *Introduction to information retrieval*, Vol. 39. Cambridge University Press Cambridge.

[34] Hládek, D., Staš, J., and Pleva, M., 2020. "Survey of automatic spelling correction". *Electronics,* ***9***(10), p. 1670.

[35] Ramya, R., and Venugopal, K., 2016. "Feature extraction and duplicate detection for text mining: A survey". *Global Journal of Computer Science and Technology,* ***16***(C5), pp. 1–20.

[36] Shelar, H., Kaur, G., Heda, N., and Agrawal, P., 2020. "Named entity recognition approaches and their comparison for custom ner model". *Science & Technology Libraries,* ***39***(3), pp. 324–337.

[37] Brownlee, J., 2018. *Better deep learning: train faster, reduce overfitting, and make better predictions*. Machine Learning Mastery.

[38] Wasserman, S., and Faust, K., 1994. *Social network analysis: Methods and applications*, Vol. 8. Cambridge university press.

[39] Freeman, L., 2004. "The development of social network analysis". *A Study in the Sociology of Science,* ***1***(687), pp. 159–167.

[40] Xiao, Y., and Sha, Z., 2022. "Robust design of complex socio-technical systems against seasonal effects: a network motif-based approach". *Design Science,* ***8***, p. e2.

[41] Cui, Y., Ahmed, F., Sha, Z., Wang, L., Fu, Y., Contractor, N., and Chen, W., 2022. "A weighted statistical network modeling approach to product competition analysis". *Complexity,* ***2022***, pp. 1–16.

[42] Sha, Z., Huang, Y., Fu, J. S., Wang, M., Fu, Y., Contractor, N., and Chen, W., 2018. "A network-based approach to modeling and predicting product coconsideration relations". *Complexity,* ***2018***, pp. 1–14.

[43] Wang, M., Sha, Z., Huang, Y., Contractor, N., Fu, Y., and Chen, W., 2016. "Forecasting technological impacts on customers' co-consideration behaviors: a data-driven network analysis approach". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 50107, American Society of Mechanical Engineers, p. V02AT03A040.

[44] Tao, K., Abel, F., Hauff, C., Houben, G.-J., and Gadiraju, U., 2013. "Groundhog day: near-duplicate detection on twitter". In Proceedings of the 22nd international conference on World Wide Web, pp. 1273–1284.

[45] Strauss, B., Toma, B., Ritter, A., De Marneffe, M.-C., and Xu, W., 2016. "Results of the wnut16 named entity recog-

nition shared task". In Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT), pp. 138–144.
[46] Rahutomo, F., Kitasuka, T., and Aritsugi, M., 2012. "Semantic cosine similarity". In The 7th international student conference on advanced science and technology ICAST, Vol. 4, p. 1.
[47] Cars.com, 2016. What were the best-selling cars in 2016? https://www.cars.com/articles/what-were-the-best-selling-cars-in-2016-1420692870639/. Accessed on March 12, 2023.
[48] Cars.com, 2017. What were the best-selling cars in 2017? https://www.cars.com/articles/what-were-the-best-selling-cars-in-2017-1420698582900/. Accessed on March 12, 2023.
[49] Farrell, J., and Rabin, M., 1996. "Cheap talk". *Journal of Economic perspectives,* ***10***(3), pp. 103–118.
[50] Chen, W., Hoyle, C., and Wassenaar, H. J., 2013. *Decision-based design: Integrating consumer preferences into engineering design*. Springer.
[51] Sha, Z., Cui, Y., Xiao, Y., Stathopoulos, A., Contractor, N., Fu, Y., and Chen, W., 2023. "A network-based discrete choice model for decision-based design". *Design Science,* ***9***, p. e7.