It's Shotime: Social Media User Identity and Engagement with Shohei Ohtani Instagram Content

Gabrielle Park[a]

This manuscript was compiled on November 27, 2024

Introduction

Over the past two decades, social media has become a significant part of our social landscape and has shaped interactions between people. Since its emergence in the early 2000s, it has brought new opportunities—along with challenges—for interactions between individuals, shaping its own rules for what drives human behavior (1). As the hybrid of technology and real-time communication, its utilization of multi-media platforms and patterns of social engagement are dictated by both computer algorithms and human interest (1). Using the metric of likes, views, and comments, social sciences and computational sciences overlap in an attempt to understand social media engagement. Studies on social media engagement have found that virality, or high-volume sharing of online content, is very human. People share and post what they feel and gravitate toward content that evokes emotions, whether joy, surprise, or anger (2). The popularity of this content can also be driven by other, technology-focused factors: the medium of delivery, the timing of posting, the platform of a post, and the use of familiar celebrities or figures (2).

This interplay between human interest and social media strategy highlights the complexity of social media. In many ways, it mimics human behavior and pre-existing identities in predictable ways: people who love baseball are likely to share social media posts about baseball and people who love dogs will likely enjoy content that features cute dogs. It can also be tied to cultural and racial identity. Social media, with its interactive and participatory nature, may be conducive to allowing one to explore their racial identity (3). At the same time, negative aspects of human behavior, such as racism and prejudice, may be replicated online in "interesting, sometimes disturbing ways" (4). Therefore, this study investigates the role that identity can play in social media engagement.

This study features Shohei Ohtani as an interesting case study. The Japanese baseball player, who currently pitchers for the Los Angeles Dodgers, has grown in prominence over the past several years—frequently the source of virality on social media. Ohtani has won multiple unanimous National League Most Valuable Player (MVP) Awards, was the first Japanese-born player to win a league home run title, and most recently, won the 2024 Major League Baseball World Series with the Dodgers (5). This past season has been particularly notable, as he had signed a 10-year, $700 million contract with the Dodgers, setting the record for the largest contract in professional sports history (5). In addition to athletic achievements, Ohtani is a fascinating figure to study because of his worldwide fan base. In addition to loyal fans from his home country Japan, he has garnered attention from baseball lovers from around the world (6). These fans often hail from countries that have been historically under Japanese influence and rule, such as Taiwan and Korea (7). Despite historic tensions with Japan, even South Koreans don his jersey and root for him, citing his kindness to Korean fans, respect for Korean national teams, and charisma (8). With Ohtani as a generally well-liked player and his unique background for a MLB pro-baseball player, this study aims to investigate what factors drive the virality of his social media posts, specifically investigating Instagram posts that reference Ohtani.. Furthermore, by stratifying based on Instagram users' content language and ethnic-identifier in usernames, I analyze what topics drive online fan interaction between users from Asian countries and users of Asian-descent.

Research Question. How does ethnicity and sub-group identity (Japanese, Taiwanese/Chinese, Korean vs. Japanese-American, Taiwanese/Chinese-American,

Significance Statement

This study explores the relationship between social media, identity, and representation. Through topic analysis comparisons between different identity groups and sub-identity group, I analyze nuances in social media engagement. I consider the capacity for machine learning techniques, such as OpenAI's API, to identify social media user identity.

Author affiliations: [a]Dartmouth College

Korean-American) affect engagement with and the content of social media posts about Japanese baseball player Shohei Ohtani?

Literature Review

Social media has grown to prominence as a powerful tool for communication, socialization, and knowledge sharing (1). While it replicates patterns of human behavior from face-to-face interactions, the dramatic development of the social media environment has shaped people's interactions. For example, people can join virtual communities and can "friend" or "de-friend" people with a click of a button (1). It provides a collaborative space that enables participation through user-generated content and information sharing (1).

Media sharing sites, such as Facebook, Instagram, and YouTube, allow users to upload and share multi-media content. While they were originally developed for individuals to share personal information and materials with friends and other people, they have evolved to become communication channels for companies to advertise products or for public figures to promote information (1). Tellis et al. investigate what drives video ad sharing across multiple social media platforms, considering how marketers and business can enhance exposure and sharing (2). Online digital content allows consumers to easily share what they like with others, exponentially increasing the total number of views. Each person who encounters media judges whether or not to consume it and whether or not to share it. Therefore, they investigate how emotional, informational, and commercial content might affect sharing (2). They find that content that evokes positive emotions, specifically inspiration, warmth, amusement, and excitement, generates more shares. They do not find negative emotions to be significant in their analysis. They also find that elements of drama, such as surprise, likable characters, and plot, elicit positive emotions and induce sharing (2). While this study focuses specifically on ad sharing and marketing, similar logic can translate to other social media where positive emotion or dramatic emotions can drive engagement. Social media is also fascinating in its capacity to replicate individual identity. A profile can serve as one's online representation of themselves and can improve their online experiences, especially with advertising (9). Each platform, depending on its purpose, can produce identity in different ways. For instance, Facebook is focused on personal self-presentation while LinkedIn's purpose is professional self-promotion and network building (9).

Social media, through its capacity to allow for self-expression can also produce identity struggles. Users can engage with interactive and participatory content as a means of exploring racial identity in positive ways (3). As people move physcially across national boundaries, online spaces of connection based on racial identity can be a positive mechanism for building sustained imagined communities (4). It can create spaces for intimate discursive interactions and can be a site for identity construction. Representation is also an important feature of social media. Content can provide individuals, especially younger social media users, with glimpses of their "possible selves" (3). Positive representations of racial identities can reflect back on young people, either inspiring them to strengthen their relationship to their racial identities or discouraging them from fully expressing their identities online (3). Evidently, the virtual nature of the internet has not taken away the negative aspects of the lived experience of race and racism in the real world (3). Especially due to social media's *social* nature, racism can be replicated online.

This study specifically focuses on baseball player Shohei Ohtani and the role that users' identities can play into their engagement and production of content. As a Japanese-born player, Ohtani brings in a large and deeply devoted fan base from Japan (6). Other East Asian countries also have a deep interest in baseball and have displayed an interest in Ohtani (8). In particular, Taiwan and Korea are known for their interest in baseball. This is often seen as a product of Japanese colonization in these countries and the consequent cultural influences (7). While this historic connection explains Asian fan interest overseas, it raises the question of how historic ties might affect Asian sub-groups in the United States. The term "Asian American," which broadly refers to Americans of Asian descent, encompasses a range of ethnic identities, familial histories, and personal experiences. From the first Chinese migrant workers during the Gold Rush of 1848-49 to the multi-generational Asian American communities today, the idea of "Asian America" has evolved over

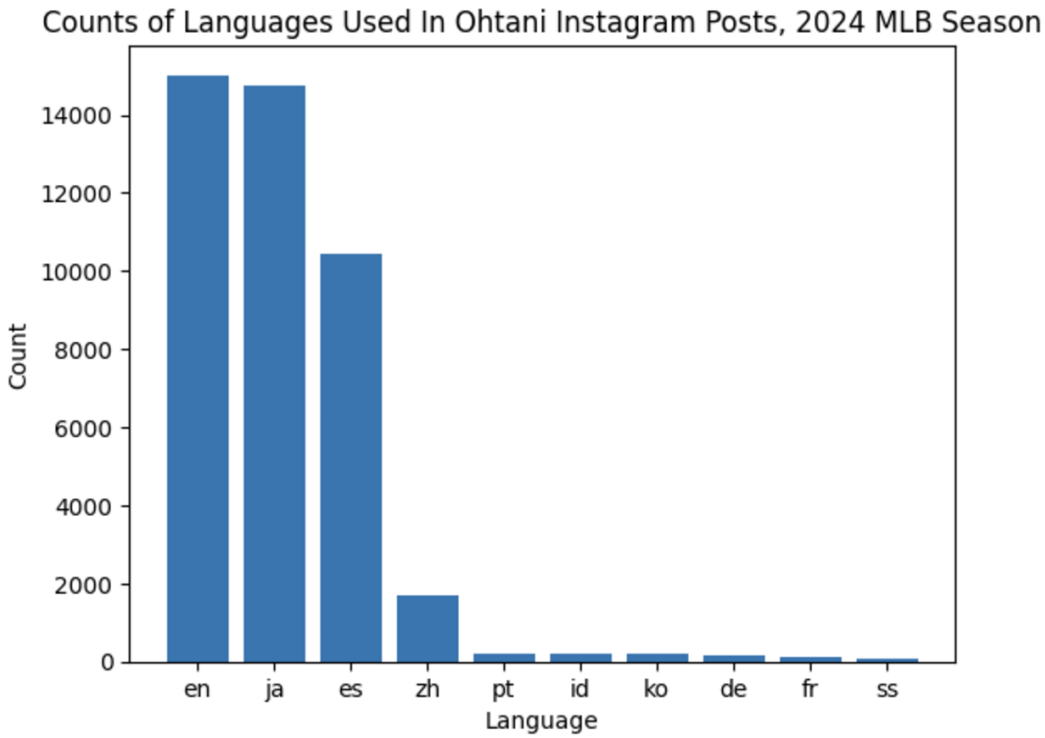

Fig. 1. Distribution of Languages in Social Media Posts

time (10). As 22 million Asian Americans trace their roots to more than 20 countries in East and Southeast Asia and the Indian subcontinent, it would be remiss to assume they all share the same experience (11). There is no such thing as one "Asian American" experience. While acknowledging this nuance, unifying within the "Asian American" identity can be a powerful opportunity to galvanize support from different ethnic subgroups, crossing ethnic and class divisions (10). While little research has studied the role of Asian/Asian-American identity in media, representation for users of Asian descent can be a driving factor in shaping racial identities in an inspirational way (3). This study hopes to bring together past research on social media and identity to understand what drives content creation and engagement from users of different racial subgroups. Using Shohei Ohtani-related social media posts as a case study, I hope to fill the gap in research on Asian and Asian sub-group social media presence.

Data

Data Collection. The data for this study comes from the Meta Content Library, which is an archive of all public posts across Facebook, Instagram, and Threads. To study social media engagement on Shohei Ohtani, social media posts were queried if the text content contained mentioned to "Shohei Ohtani" across different languages. It was scraped from the past couple years and includes metadata and engagement data.

For the purposes of this study, I use data on Instagram posts because the nature of the platform encourages more people to post publicly and engage with content from people they may not necessarily follow. Facebook was considered as a potential data source for this project, especially with the accessibility of data on users' first and last names. In preliminary analyses, I found that Facebook tended to display content from businesses rather than personal accounts, so I opted to use Instagram data.

Data Cleaning. The raw dataset included 69,042 rows and 18 columns. Columns included data on content type (medium of delivery), creation time, user identifiers, engagement statistics (likes, comments, views), and text. For this study, I select content type, hashtags, language, post owner type, number of likes, and text.

From the almost 70,000 rows, the most represented languages are English, Japanese, Spanish and Chinese. See Figure 1.

The time frame selected for the dataset is from November 2023 to November 2024, in order to capture the full 2024 MLB season and any news about Ohtani signing his new contract with the Dodgers at the end of 2023. See Figure 2.

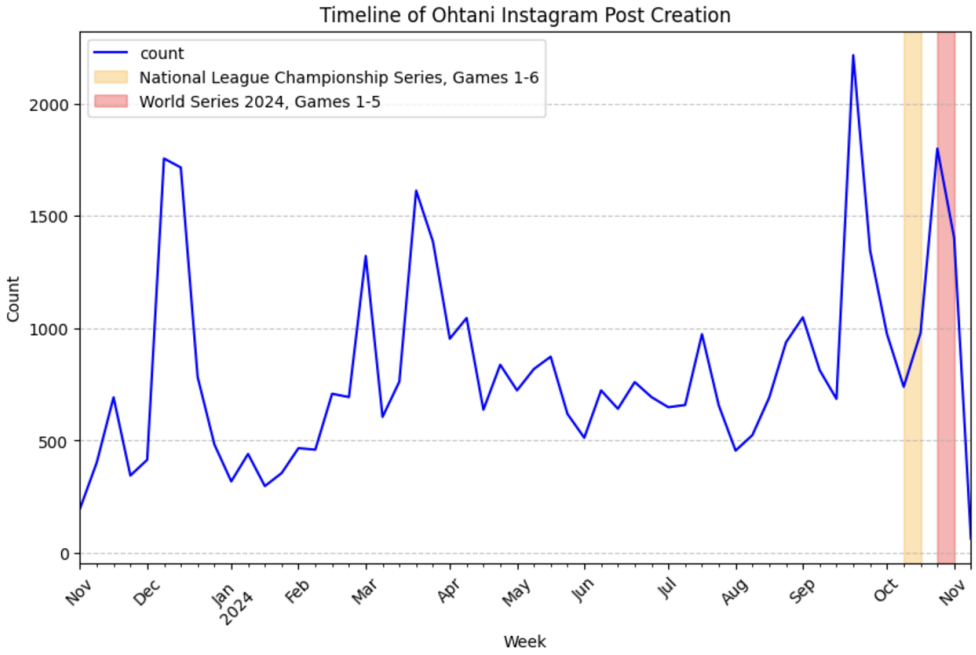

Fig. 2. Timeline of Ohtani Instagram Post Creation

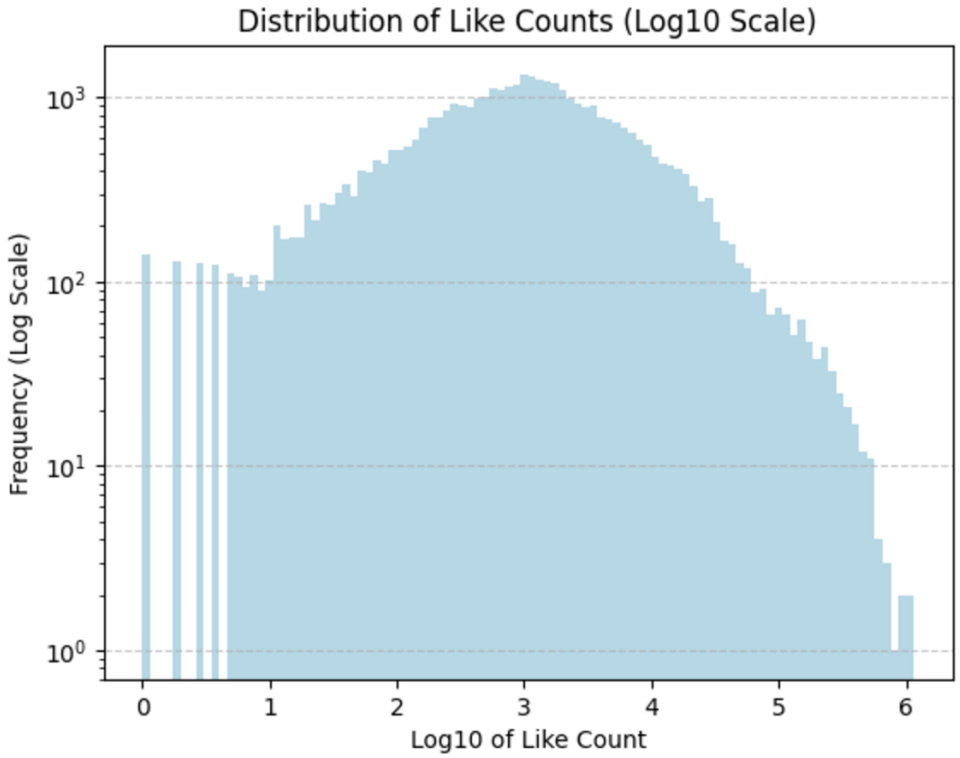

Fig. 3. Distribution of Like Counts in Ohtani Instagram Posts

Then, I select posts from only creator and personal account, which are more likely to be from individuals. While business data would provide interesting insights into how accounts market to users using sports figures like Ohtani, this study would benefit most from content produced by individual users.

For my analyses, I clean the data. I encode the categorical labels to be numeric, specifically post owner type and content type. I apply a base-10 logarithmic transformation to the like counts, since there is a large range between posts with few likes and posts with many likes. See Figure 3. Then, I count up the number of hashtags used and drop posts with NAs in significant columns, such as the number of likes.

Data Processing. A significant aspect of this project was the use of OpenAI's API, which is an interface to access OpenAI's AI models to analyze and generate content, similar to other Large Language Models (LLMs).

I use this tool to analyze user name and text data. This is helpful for deriving post owners' likely ethnic/cultural background and for translating large amounts of text data. For this project, I create large batches of API requests for processing, using GPT-3.5 Turbo. First, I analyze post owners' name data, which is the username (e.g. @ShoheiOhtani) and their name (e.g. Shohei Ohtani). While the dataset provides text language data, which easily allows us to infer where a user comes from, there is no data on English-speaking users' racial background. Therefore, by inputting the user's name data into OpenAI's models, it can infer name etymology and likely sub-group identity. For example, while 'Josh Kim' might be an American and post content in English, the last name Kim can clue us into his sub-group identity as Korean American. The prompt used for this processing is "Analyze username information to infer: "name_origin": the user's identity or likely ethnic background (0: Chinese-American, 1: Korean-American, 2: Japanese-American, 3: Other Asian-American, 4: Hispanic/Latino-American, 5: Other, 6: Unsure). Output as JSON." Note: there are limitations for this approach, which are considered in the Discussion section. Then, I follow a similar approach to translate non-English and non-Spanish language text data into English. Once the data processed, I join it back with the cleaned data frame to be used in analyses.

Once the text data is translated to English, I run a sentiment analysis of the text using VADER (Valence Aware Dictionary and sEntiment Reasoner) and save the compound sentiment score. It ranges from -1 to 1 with -1 indicating negative sentiment, 0 indicating neutral, and 1 indicating positive.

Methodology

To analyze differences in social media content between identity groups and sub-groups, I run Latent Dirichlet Allocation (LDA) models. This is a machine learning technique in natural language processing to estimate the distribution of words over topics and the proportion of documents in each topic. I run LDA analyses for the whole data set and then separately for the identity groups of interest to highlight differences in social media content about Ohtani. I use the Gensim Python library to visualize the topic modeling.

Lastly, I train a CatBoost regressor model to predict the number of likes a post receives using post features (i.e. content type, post owner type, Vader sentiment analysis score, hashtag counts, and language). I use SHAP (SHapley Additive exPlanations) to interpret the output of the CatBoost model. This is helpful for determining the importance of each feature in the model's prediction.

Results

When running a LDA model on all of the text data, I find that most of the topics are similar to each other, as they all address Ohtani and baseball. See Figure 7. Despite this, there are some nuanced differences. Topic 0 references Shohei Ohtani, his team the Dodgers, and contract, indicating that it relates to his recent contract with the Dodgers. Other topics reference him, his team, baseball-terms, and his achievements. See Figure 4. Only Topic 4 seems to be unrelated to baseball, referencing terms like "flower," "photograph," and "rose." The output from the LDA for Topic 4, including the most important terms is pasted below: (4, 0.015*"flower" + 0.012*"photograph" + 0.010*"ranger" + 0.008*"rose" + 0.008*"texa" + 0.006*"surround" + 0.006*"photographi" + 0.005*"pensarnocuesta" + 0.005*"laughter"') Subsetting for Japanese language posts, there is a similar focus on "shohei", "ohtani", "shoheiohtani", and "dodger", "angel", and "player." This indicates that it focuses on the player and his relation to his team. Topic 2 features Japanese-language terms and his fan base, indicating an interest in his fan base or fan-related content. Other topics generally address media coverage baseball. Topic 5 similarly references "flower", "photograph", and "rose" and is unrelated to baseball.

Korean language posts focus more on Los Angeles, the home of his team the Dodgers, and tourism ("tour inquiries", "Los Angeles tourism", "overseas travel", etc.). This could be due to the large Korean population in Los Angeles. There is also references in Topic 2 to figures in K-Pop and BTS, indicating that there may be crossovers in Ohtani and K-Pop fanbases.

Then I focus on Chinese language posts. While Meta does not make a distinction between Simplified and Traditional Chinese, most of the posts were written in Traditional Chinese, which is spoken in Taiwan. Topics in Chinese focus more on Ohtani's athletic performance and status. For example, Topic 2 references "game", "season", "look", and "DAZN" (a streaming platform). This demonstrates there may be more interest in watching baseball as a game and in Ohtani as a player.

To consider the nuance of Asian/Asian-American identity, I also run topic analyses based on which English-speaking users were likely of Asian descent. Among this population, the topics had many overlaps, mostly focusing on terms like "stadium," "train," "player," "base" demonstrating an overall interest toward the Dodgers and Ohtani's performance.

Users who were identified as likely Japanese-American referenced topics that were overall interested in "stadium," "park," "player," and "field," which is not very distinct from English-speakers of Asian descent.

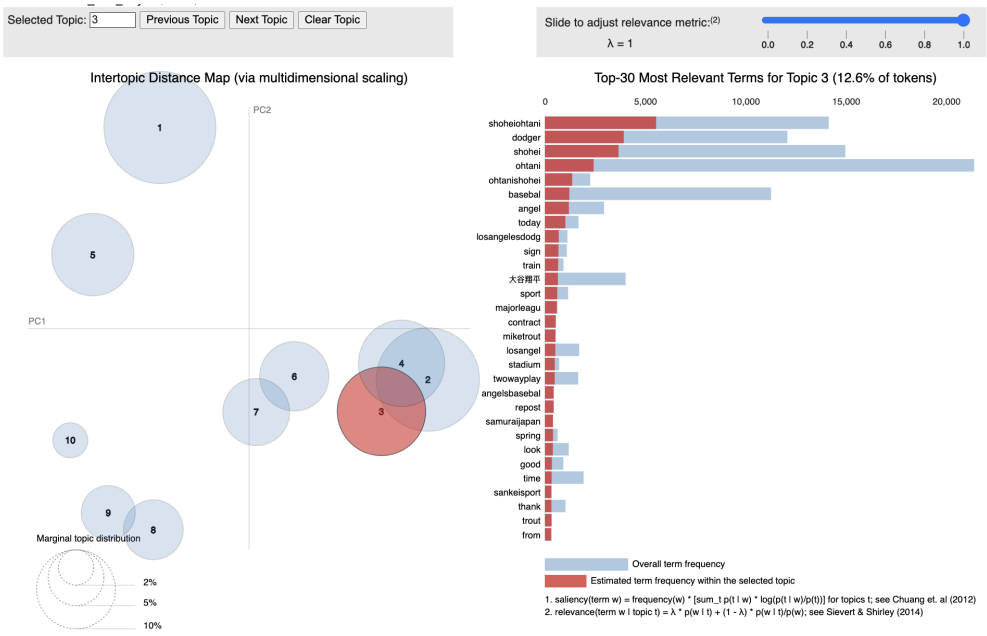

Fig. 4. LDA Model for All Posts

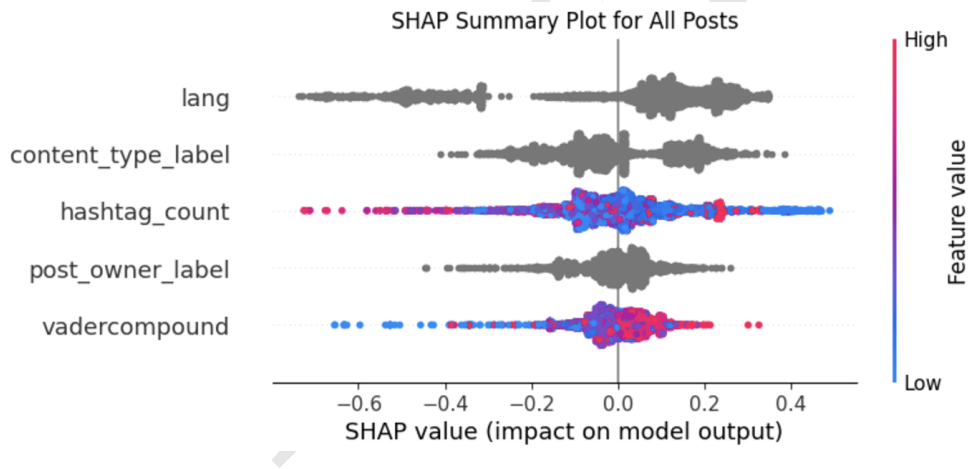

Fig. 5. SHAP Summary Plot for All Posts

Users who were identified as likely Korean-American referenced topics that were more distinct. For instance, Topic 1 discussed Shohei, Dodger, Ohtani, Shohei Ohtani, Lancia, smile, Soohoo, that, scrapbook, Godzilla, suggesting references to social media posts. Topic 3 referenced Bangkok, place, Bambam, Louis Vuitton, Shohei, exhibit, journey, maison, Ploenchit, Gaysorn, which could relate to Ohtani's brand collaborations and off-field interactions. This is distinct from posts by Korean-speaking users.

Lastly, looking at English speaking users of Chinese-descent, topics focused on non-game time interactions. Japanese food, Japan, and cities in Japan are referenced, whereas this had not been referenced as much in other subgroups.

Overall, these results are not as distinct, as they all reference Ohtani and as a result reference baseball. Still, there are nuanced differences in the types of topics mentioned by each group.

The CatBoost regressor model did not produce significant results in the data set overall and for each sub-group, potentially due to the limited data. Overall, the model including all Instagram posts has an R-squared value of 0.263, which indicates that the model is not predicting a large proportion of the variance in the data. Still, by visualizing in a SHAP Summary plot, it is evident that high positive sentiment has a greater impact on the model, and lower sentiment is not as significant. Also, low hashtag counts has a greater impact on the model, and high hashtag counts has a negative impact. See Figure 5.

For Japanese-language posts, the Cat Boost regressor model is greater at predicting variability with an R-squared of 0.326, but this is still a low value. Similarly, we see that positive sentiment is more influential. Also, low hashtag counts have a greater impact on the model output. See Figure 6.

Other English-speaking sub-groups did not produce significant enough results, potentially due to the low number of posts.

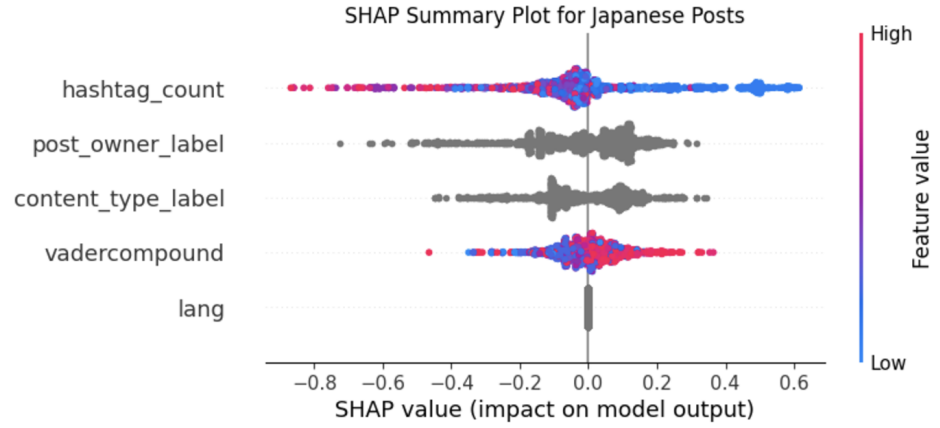

Fig. 6. SHAP Summary Plot for Japanese-language Posts

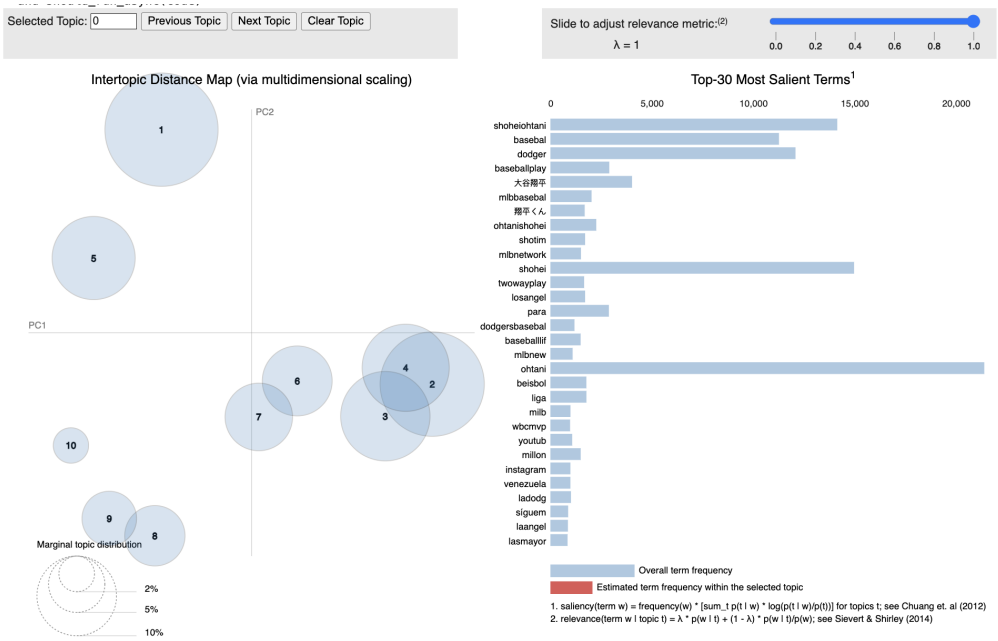

Fig. 7. LDA Topic Model for All Ohtani Posts

Discussion and Conclusion

This study explores the relationship between social media, identity, and representation. Through topic analysis comparisons between different identity groups and sub-identity group, we see that engagement differs. While most posts related to Ohtani, baseball playing, and the Dodgers, we see some nuanced differences. For instance, Japanese-language posts mentioned his Japanese fanbase, Korean-language posts focus on Los Angeles and tourism, and Chinese-language posts focus on game play. Looking at Asian-American identity, we do not find distinctive results, but we do see that topics between Asian-identity subgroups and their respective language-speaking groups are not the same.

A main limitation of this study is the capacity to identify users based on their social media presence. While social media is a replication of identity online and can be a positive force in building community and exploring racial identity, it must be acknowledged that clear assumptions cannot be made about one's identity based on their curated social media presence (3).

Furthermore, with the use of OpenAI's API, it raises the question of how well identity can be inferred from names. There is much ambiguity in names, espcially if users withhold information for privacy reasons. In addition, there is limits in machine learning models' ability to understand the nuances in identity. For example, someone's family name could be American, but they could have a cultural background that is not referenced in their name. Also there are family names that are ambiguous between cultural groups, such as Lee. Also, with translation software, there are limitations in multilingual large language models' capability to analyze sentiment and nuance in foreign languages with nuance and accuracy. While this study uses OpenAI's API to translate non-English and non-Spanish text, there is a need for other resources to translate and analyze foreign natural language.

These limitations are important to consider and can point to areas of research that should be further explored in the future. One's identity and how they engage with it are complicated subjects. The prevalence of social media adds further complexity to the lived experience and self-representation of identity, but it creates an opportunity to use innovative computational tools to analyze data. Future studies should consider these nuances and push for more research to better understand how sub-group identity can create differences in social media engagement with viral topics like Shohei Ohtani and other prominent figures in media.

Note: ChatGPT was used in this project for debugging OpenAI API batch processing code.

1. EW Ngai, KM Ka-leung, SS Lam, ESK Chin, SS Tao, Social media models, technologies, and applications. **115**, 769–802 (year?).
2. GJ Tellis, DJ MacInnis, S Tirunillai, Y Zhang, What Drives Virality (Sharing) of Online Digital Content? The Critical Role of Information, Emotion, and Brand Prominence. **83**, 1–20 (year?).
3. J Chan, Racial Identity in Online Spaces: Social Media's Impact on Students of Color. **54**, 163–174 (year?).
4. J Daniels, Race and racism in Internet Studies: A review and critique. **15**, 695–719 (year?).
5. Shohei Ohtani (year?).
6. S Wade, Fans in Shohei Ohtani's hometown gather early to watch World Series — AP News (year?).
7. SA Hargrove, Baseball and Politics in East Asia (year?).
8. Hj Kim, Ohtani may be South Korea's most-beloved Japanese athlete. His charm is winning over historic rivals (year?).
9. R Zafarani, L Tang, H Liu, User Identification Across Social Media. **10**, 16:1–16:30 (year?).
10. AP Kambhampaty, AAPI History: Activist Origins of the Term 'Asian American' — TIME (year?).
11. NG Ruiz, A Budiman, Key facts about Asian Americans, a diverse and growing population (year?).