Japanese Beauty Marketing on Social Media: Critical Discourse Analysis Meets NLP

Emily Öhman and **Amy Grace Metcalfe**[*]
Waseda University
ohman@waseda.jp and agmetcalfe@toki.waseda.jp

Abstract

This project is a pilot study intending to combine traditional corpus linguistics, Natural Language Processing, critical discourse analysis, and digital humanities to gain an up-to-date understanding of how beauty is being marketed on social media, specifically Instagram, to followers. We use topic modeling combined with critical discourse analysis and NLP tools for insights into the "Japanese Beauty Myth" and show an overview of the dataset that we make publicly available.

1 Introduction

Instagram is one of the most widely used social media platforms in Japan, with an audience of over 38 million (Clement, 2020). Instagram's focus on photo sharing generates a saturation of images conveying beauty. Since Naomi Wolf's 1991 publication "The Beauty Myth" (Wolf, 1991), countless studies have shown the negative effects the beauty industry has on women, including obsessions over weight (Kayano et al., 2008), anxiety issues (Miller, 2006), and furthering inequality between the sexes (Walter, 2011).

This pilot study examines the linguistic features of Instagram posts by beauty companies, both qualitatively and quantitatively, in order to discover how language contributes to the construction of beauty standards in the Japanese context. We use established natural language processing (NLP) methods and topic modeling, combined with critical discourse analysis (CDA) to understand both the quantifiable data and the social effects of these Instagram posts. This study contributes to a better understanding of the definition of beauty within the Japanese context and offers additional insights into beauty ideals and how these are fabricated. Furthermore, the dataset is made public and will therefore be of use to other researchers as well.

We have chosen to aim our attention at posts made by make up, skincare, hair removal, and hair care companies for the purposes of this study. The reasoning for this choice is due to the the "opt-in" nature of these practices. While other practices such as fashion, or dental care could be considered beautification practices, we have chosen to omit these, as they are either culturally required in the case of clothing (Rouse, 2017) or done primarily for hygiene purposes in the case of dental care. Thus, the companies we have chosen focus only promote practices which are not required for hygiene or protection of modesty.

In the following sections we present the background for our study, explain our data and methods, examine the results and analyze and discuss them in a wider context. We conclude with a discussion including future work related to the dataset.

2 Background

The objectivity of qualitative analyses have been criticized for being too subjective (Cheng et al., 2013). Whether this is entirely justified or not (see e.g. Baškarada and Koronios (2018)), qualitative analyses supported by quantitative methods have proven useful in investigating media discourse, allowing researchers to identify textual patterns (O'Halloran, 2010). Incorporating NLP and topic modeling provides scaffolding for the CDA, leading to verifiable empirical results.

Asian beauty trends have been researched in a multitude of different fields, including but not limited to medicine (Liew et al., 2016), marketing (Li et al., 2008), and sociology and gender studies (Saraswati, 2020). Asia is a diverse market in terms of consumer demographics, culture, and beauty ideals (Yip et al., 2019; Yip, 2018). However, some

[0] Both authors contributed equally to this paper.

Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 131–137
co-located with ICON-2021, December 16-19, 2021, Silchar, India. ©2021 NLP Association of India (NLPAI)
https://doi.org/10.26615/978-952-94-5833-2_015

commonalities can be seen in terms of beauty ideals such as idealization of whiteness and a desire for double eye-lids (Saraswati, 2020).

Although the concept of "whiteness" is not only about idealizing Caucausian skin types, and is traditionally linked to socio-economic status (SES) in Asia, most scholars agree that Western standards for female beauty ideals are prevalent in Asia and include "whiteness" (Jung, 2018).

Surprisingly little work has been done on the topic of beauty by using computational methods. Gender and online slurs have been explored in many papers with the help of NLP tools, but to the best of our knowledge there has not been a study on beauty ideals using methods from even corpus linguistics on modern data. The closest thing we could find was a qualitative study comparing the ideals of female beauty in Malaysian and Belgian advertising (De Cort, 2009), but this study used no NLP methods, and only partially relied on corpora.

In discourse analysis, the topic is well-explored, but relying mostly on qualitative analyses and in some cases linguistic features, Asian beauty ideals are also a common topic (Iqbal et al., 2014; Xu and Tan, 2020; McLoughlin, 2017; Renaldo, 2017).

A related topic is the gendered choice of script in Japanese (Wakabayashi, 2016; Maree, 2013) which is both linked to the producer of the text and the intended audience (Iwahara et al., 2003; Dahlberg-Dodd, 2020; Mellor, 2003). Japanese can be written with three different domestic(-ish) scripts (kanji, hiragana, and katakana). Kanji is usually used for content words, hiragana for syntactic particles and similar but can also be seen as cute and girly (Robertson, 2019), and katakana is used for loan words and emphasis. Additionally the Latin alphabet is also used. Additionally, 和語 (native Japanese words) 漢語 (Sino-Japanese words) and 外来語 (loanwords) can be used to achieve different effects in a similar way to script-choice.

Primarily Fairclough's three-part model (Fairclough, 2001) has been used as a theoretical framework for critical discourse analysis of advertising (see e.g. Kaur et al. (2013) and Lestari (2020)). Consisting of the micro, meso, and macro levels of interpretation, it allows for a comprehensive analysis of texts. At the micro level, linguistic features are identified and investigated. At the meso level, we can see the strategies used and how the message is conveyed, and finally the social context and social effects of such texts will be considered via the macro level. Fairclough's model is particularly well-suited when analyzing social and cultural change, due to the comprehensive nature of the framework (Iqbal et al., 2014).

3 Data

The data was collected using the Instaloader[1] package for Python. The Instagram profiles of twenty companies were randomly chosen, with the criteria being that the company's product or service must be available in Japan and that they have an Instagram profile aimed at the Japanese market. In the cases where the brand had many profiles, the profile mainly written in Japanese was selected. The companies chosen were; Beauteen (beauteen_offical), Chifure (chifure_official), Curél (curel_official_jp), DHC (dhc_official_jp), Etude House (etudejapan), Ichikami (ichikami_kracie) , Innisfree (innisfreejapan), Kanebo (kaneboofficial), Kireimo (kireimo_official), Kosé (kose_official), Liese (liese_official_jp), Maybelline (maybellinejp), Musée (museeplatinum_insta), Palty (paty_official), Revlon (revlonjapan), Rimmel (rimmellondon_jp), Rize Clinic (rizeclinic), Sekkisei (sekkisei.official), Shiseido (shiseido_japan), and TBC Aesthetic (tbc_aesthetic).

This resulted in 7477 posts by these twenty companies, for a total of 365752 lemmas. Depending on the method of adjective inclusion, between 8% and 17% of these were adjectives. The dates of these posts ranged from early 2016 to June 2021. We focused on the text content of the posts and this data is freely available on GitHub[2].

4 Methods

The json-formatted data was then converted to pandas dataframes and flattened for exploratory data analysis. The posts themselves were segmented, tokenized, lemmatized and annotated for part of speech using Fugashi (McCann, 2020), mecab (Kudo, 2006), and spaCy (Honnibal and Montani, 2017). We also used the medium sized Japanese language model for spaCy to improve word recognition.

This mix of general tools and specific tools for Japanese gave us access to both the Western speech tags, such as ADJ - adjectives, but also Japanese tags which made it possible to include 形容詞

[1] https://instaloader.github.io/

[2] https://github.com/esohman/JapaneseBeauty1

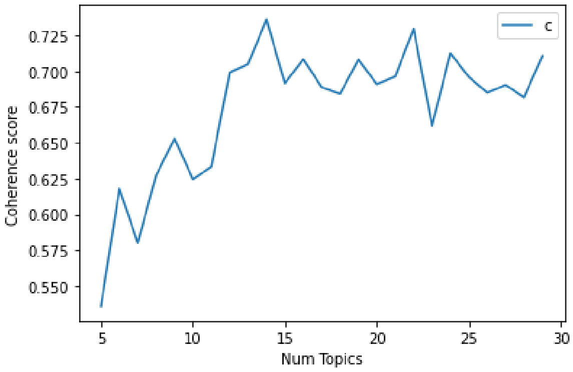

Figure 1: Coherence Values

(adjectives) as well as 形容詞可能 (nominalized adjectives and adjectivized nouns) in our keywords, and therefore not miss important keywords due to overly general linguistic features and part of speech categorization. We chose to focus on adjectives for the linguistic analysis as a quick access point to descriptive language.

The raw text was then cleaned up and used for LDA-based topic modeling using Latent Dirichlet Allocation (LDA) with Mallet as well as NMF-based. It might have made sense to try 20 topics as we had 20 companies, but we assumed many of these companies would be posting somewhat similar content on similar topics and therefore wanted to empirically ascertain the ideal number of topics. We used coherency scores to find the most suitable number of topics for our dataset (see table 1). After 12 topics the coherence scores to sharply rise, topping at 14 with a coherence score over 0.73.

As the dataset is still quite small, we experimented empirically with varying numbers of topics with a few different models (LDA, LDA with Mallet, NMF with Kullback-Leibler and NMF with Frobenius norm). We found that for this data, our most human-interpretable results were achieved by choosing 14 topics with the NMF model using generalized Kullback-Leibler divergence (see figure 2). However, the effect of the different segmentation (spaCy, mecab) and the inclusion of trigrams and bigrams, as well as the use of tfidf vectorizer made the results of these models differ in minor ways but with an impact on the final results. Therefore both the LDA with Mallet model and the NMF with Kullback-Leibler were used for the final analysis with the NMF model getting the clearest and easiest to interpret topics, but with the LDA model finding some interesting underlying topics that were not similarly present in the NMF model.

5 Results

Some of the most representative topics came from the LDA with Mallet model and include those in table 1.

Although the NMF models[3] seemed to be better at homing in on the product types, the LDA models seemed better at finding underlying topics such as mask makeup and self-care, as well as different types of skincare as their own topics.

Another common incursion into the most frequent tokens in topics was brand names. Usernames were stripped from the text, so these occurrences were cases of self-promotion.

We also looked at spelling differences between these adjectives. We looked at the word *kirei* (beautiful, clean) in particular. There were 880 instances of *kirei* in the data of which 52 were of Latin spelling (romaji), 147 were kanji, 83 were hiragana, and 598 were katakana, which also seemed to be favored in hashtags for other words as well.

As can be seen in figure 3, the adjectives and adjectival words have an emphasis on the sensory experience of the products, i.e. how will they make the consumer's skin feel or look like. *Soft, moist, supple, fluffy, smooth, velvety, glowy, dewy* and similar concepts appear alongside *cute* and *just-right*. With the most common adjective(-like) concept being *like/love*.

6 Analysis & Discussion

Overall, patterns that emerge from the corpus, reveal elements of what beauty means in the Japanese context, namely the inclusion of brand equity, and the importance of the skin.The scientific basis for the formulation of components of products seems to also be becoming more common as can be seen in one of the topics (special skincare) in table 1.

Some of the topics that emerged were related to self-care in general (topic 10 in the NMF model). Another topic that emerged from the LDA model was related to COVID-19 and self-care and included words like "mask makeup" (258 occurrences). Most topics were, however, clearly related to cosmetics, skincare, or hair care and included words describing creams, colors, essences, skin tone (yellow-base and blue-base, similar to the terms commonly used in English warm and cool toned). "Whiteness" was a word that came up in

[3] The undecoded characters in Topic 8 are Korean words. We could not find a way to render Japanese and Korean characters simultaneously in a matplotlib plot.

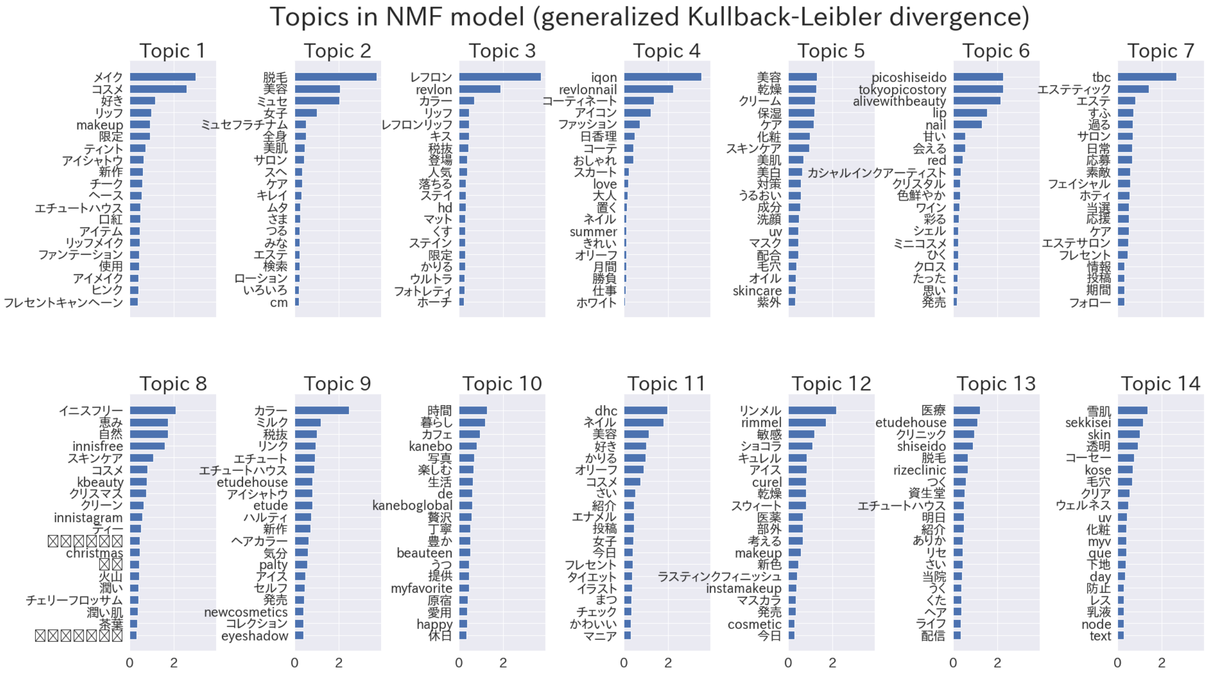

Figure 2: Topics of 14-topic NMF model using Kullback-Leibler divergence

beauty, skin	EN	special skincare	EN	foundation	EN
化粧	make-up	肌	skin	感	feeling
ケア	care	乾燥	dry	毛穴	pore
スキンケア	skin care	敏感	sensitive	肌	skin
クリーム	cream	成分	component	ファンデーション	foundation
美肌	beautiful skin	保湿	moisturizing	ベース	base (loanword)
肌	skin	配合	formulation	下地	base (native JP)
美白	whitening	性	sex, gender	ツヤ肌	dewy skin
UV	UV	picoshiseido		仕上	Finish
美容	Beauty	tokyopicostory		くすみ	Dull

Table 1: Example topics from the LDA with Mallet model

several topics and was not limited to one particular topic, or even subtopic such as skincare. "Whiteness" was most present in the LDA topic, and we believe the reason might be related to the different lemmatizations and segmentation of Japanese (spaCy vs. mecab) as the NMF model seemed to segment the word 美白 as *beautiful* and *white* rather than *whitening*.

Using Fairclough's model to analyze these captions, we can see that certain linguistic features and word choices arise, including a heavy use of adjectives, references to skin issues and solutions, and references to "wellness" culture. Different companies use adjectives differently and these groupings can be seen in both the list of adjectives and adverbs used to describe these ideals as well as in the topics that emerge from topic modeling. Some of the more interesting topics include COVID-related issues such as novel words for mask-specific makeup (i.e. マスクメイク、新商品、美肌作り) and how to take care of yourself and your skin (丁寧な暮らし、暮らしを楽しむ、贅沢な時間、美容好きな人と繋がりたい) during these times.

To analyze the text on the meso level, it is clear that these brands make use of Aristotle's three persuasive strategies (Mooney and Evans, 2018), but in particular ethos, or the credibility of a text. This is achieved through the use of celebrity endorsement which was commonplace among brands. For example, the hair removal clinic TBC have had a long standing partnership with model and personality Rola. When we consider more widely the

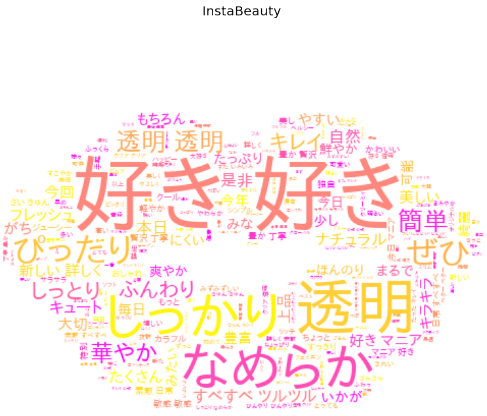

Figure 3: Visualization of adjectives and 形容詞、形容詞可能 i.e. adjectives and adjectivized nouns in the corpus

interaction these posts have with their audiences and with society, it gives the impression that these messages are portrayed as being written by someone who cares about your well being, and not a company, thus creating the illusion of friendship. This narrative, paired with the persistent collaborating between brands and celebrities, can contribute to parasocial relationships between brands and their consumers (Yuan et al., 2016).

Notably, the images in the majority of the posts we collected did not actually depict a model or a person at all, and had a greater focus on the product. This is in contrast with billboard advertisements in Japan which typically depict a human model. However, when people were pictured in these social media posts, there was a strong tendency to display only parts of the body, such as a hand, or a typically "beautiful" model (thin, able-bodied, flawless skin etc.). These are the images which may be problematic (Etcoff and Paxton, 2016). The paper explicitly states that self esteem issues increase when women view idealized advertisements, and that has increased since their 2004 report both globally and in Japan (Etcoff and Paxton, 2016). There is strong evidence to suggest that such self esteem issues have real life implications. Kayano et al. (2008) investigation further emphasizes the prominence of eating disorders, abnormal dieting, and an obsession with thinness for both genders among Japanese young people stating that young people are heavily influenced by the media (Kayano et al., 2008)).

It should be noted that the depiction of women in these advertisements differs from the depictions of men whose entire bodies are typically displayed in contrast to women who are often shown in parts. This is a known phenomenon in advertising which perpetuates a "pattern of seeing women as collections of body parts" erasing female identitites (Goldman, 2005; Schroeder and Borgerson, 1998).

7 Conclusions and Future Work

To summarize, we can see that the language and imagery used in Japanese beauty posts on Instagram do reflect modern beauty standards, and make reference to topical issues such as COVID.

We expect this corpus to yield many more interesting insights into Japanese beauty ideals, COVID-related self-care, and societal issues involving the pressure to conform to Japan-specific beauty standards. With future work we hope to dive deeper into the data and specifically we hope to add data produced by consumers themselves. It would be an interesting counterbalance to the marketing speak to be able to compare the content the average consumer outputs with the output of the companies.

Exploring the choice of script might shed more light on the study of loan-word intrusion into Japanese. As script choice is highly gendered in Japanese and there is a lot of promise in this dataset to shed further light on in what contexts the different scripts are used in social media marketing.

We plan on scraping Twitter as well as the comments of YouTube beauty-content producers, as well as personal blogs to achieve this augmentation of our dataset. It would be very interesting to look at the data from a diachronic perspective and see how COVID-19 has affected the consumers' view of beauty products.

References

Saša Baškarada and Andy Koronios. 2018. A philosophical discussion of qualitative, quantitative, and mixed methods research in social science. *Qualitative Research Journal*.

Winnie Cheng et al. 2013. Corpus-based linguistic approaches to critical discourse analysis. *The encyclopedia of applied linguistics*, pages 1353–1360.

Jessica Clement. 2020. Leading countries based on Instagram audience size as of october 2020. *Statista. https://www.statista.com/statistics/578364/countries-with-most-instagram-users/*.

Hannah E Dahlberg-Dodd. 2020. Script variation as audience design: Imagining readership and community in japanese yuri comics. *Language in Society*, 49(3):357–378.

Anne De Cort. 2009. The ideal of female beauty in two different cultures: Socio-cultural analysis of belgian and malaysian print advertisements. *Novitas-ROYAL*, 3(2).

Nancy Etcoff and Susan Paxton. 2016. The Dove global beauty and confidence report.

Norman Fairclough. 2001. *Language and power*. Pearson Education.

Robert Goldman. 2005. *Reading ads socially*. Routledge.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Asma Iqbal, Malik Haqnawaz Danish, and Maria Raja Tahir. 2014. Exploitation of women in beauty products of fair and lovely: A critical discourse analysis study. *International Journal on Studies in English Language and Literature*, 2(9):122–131.

Akihiko Iwahara, Takeshi Hatta, and Aiko Maehara. 2003. The effects of a sense of compatibility between type of script and word in written japanese. *Reading and Writing*, 16(4):377–397.

Jaehee Jung. 2018. Young women's perceptions of traditional and contemporary female beauty ideals in china. *Family and Consumer Sciences Research Journal*, 47(1):56–72.

Kuldip Kaur, Nalini Arumugam, and Norimah Mohamad Yunus. 2013. Beauty product advertisements: A critical discourse analysis. *Asian social science*, 9(3):61.

Mami Kayano, Kazuhiro Yoshiuchi, Samir Al-Adawi, Nonna Viernes, Atsu SS Dorvlo, Hiroaki Kumano, Tomifusa Kuboki, and Akira Akabayashi. 2008. Eating attitudes and body dissatisfaction in adolescents: Cross-cultural study. *Psychiatry and Clinical Neurosciences*, 62(1):17–25.

Taku Kudo. 2006. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. jp*.

Eka Marthanty Indah Lestari. 2020. A critical discourse analysis of the advertisement of Japanese beauty products. *IZUMI*, 9(1):58–74.

Eric PH Li, Hyun Jeong Min, and Russell W Belk. 2008. Skin lightening and beauty in four asian cultures. *ACR North American Advances*.

Steven Liew, Woffles TL Wu, Henry H Chan, Wilson WS Ho, Hee-Jin Kim, Greg J Goodman, Peter HL Peng, and John D Rogers. 2016. Consensus on changing trends, attitudes, and concepts of asian beauty. *Aesthetic plastic surgery*, 40(2):193–201.

Claire Maree. 2013. Writing one: Deviant orthography and heternormativity in contemporary japanese lifestyle culture. *Media International Australia*, 147(1):98–110.

Paul McCann. 2020. fugashi, a tool for tokenizing Japanese in Python. *arXiv preprint arXiv:2010.06858*.

Linda McLoughlin. 2017. *A critical discourse analysis of south asian women's magazines: Undercover beauty*. Springer.

Andrew Mellor. 2003. A survey of roman script in written japanese media. *Journal of the Faculty of Policy Management Yokkaichi University*, 2(1_2):101–117.

Laura Miller. 2006. *Beauty up: Exploring contemporary Japanese body aesthetics*. Univ of California Press.

Annabelle Mooney and Betsy Evans. 2018. *Language, society and power: An introduction*. Routledge.

Kieran O'Halloran. 2010. How to use corpus linguistics in the study of media discourse. In *The Routledge handbook of corpus linguistics*, pages 563–577. Routledge.

Zainal Renaldo. 2017. Analysis of linguistic features of beauty product advertisements in cosmopolitan magazine: A critical discourse analysis. *TELL-US Journal*, 3(2):141–54.

Wesley C Robertson. 2019. Why can't i speak in kanji?: Indexing social identities through marked script use in japanese manga. *Discourse, Context & Media*, 30:100297.

Elizabeth Rouse. 2017. Why do people wear clothes? In *Fashion Theory*, pages 122–125. Routledge.

L Ayu Saraswati. 2020. Cosmopolitan whiteness: The effects and affects of skin-whitening advertisements in a transnational women's magazine in indonesia. *Meridians*, 19(S1):363–388.

Jonathan E Schroeder and Janet L Borgerson. 1998. Marketing images of gender: A visual analysis. *Consumption, Markets and Culture*, 2(2):161–201.

Judy Wakabayashi. 2016. Script as a factor in translation. *Journal of World Literature*, 1(2):173–194.

Natasha Walter. 2011. *Living dolls: The return of sexism*. Hachette UK.

Naomi Wolf. 1991. *The beauty myth: How images of beauty are used against women*. Random House.

Huimin Xu and Yunying Tan. 2020. Can beauty advertisements empower women? a critical discourse analysis of the sk-ii's" change destiny" campaign. *Theory and Practice in Language Studies*, 10(2):176–188.

Jeaney Yip, Susan Ainsworth, and Miles Tycho Hugh. 2019. Beyond whiteness: Perspectives on the rise of the pan-asian beauty ideal. In *Race in the Marketplace*, pages 73–85. Springer.

Jesse Wai Chi Yip. 2018. Communicating social support in online self-help groups for anxiety and depression: A qualitative discourse analysis. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Chun Lin Yuan, Juran Kim, and Sang Jin Kim. 2016. Parasocial relationship effects on customer equity in the social media context. *Journal of Business Research*, 69(9):3795–3803.