Linguistic and Paralinguistic Features influencing Reliability Judgments of Thai Twitter Reviews

Nanthicha Angsuwichitkul[1], Warisaraporn Limprasert[1], Francesco Burroni[2]
[1]Faculty of Arts, Chulalongkorn University, Thailand
[2]Institute for Phonetics and Speech Processing, LMU Munich, Germany
Center of Excellence in southeast Asian Linguistics, Faculty of Arts, Chulalongkorn University
{6340114222,6340210822}@student.chula.ac.th
francesco.burroni@phonetik.uni-muenchen.de

Abstract

Due to the current widespread use of social media, deceptive opinion spam has risen, especially on Twitter, and distinguishing between real and fake reviews is a difficult task. Previous studies have highlighted the importance of using several features, including linguistic, stylometric, and others, to differentiate between truthful and deceptive opinions in English reviews. Extending previous work, we study the linguistic and paralinguistic features associated with the subjective impression of reliability of Thai-language reviews on Twitter. Our findings show that features such as the presence of URLs, hashtags, the number of exclamation marks, and the use of first-person pronouns have an effect on the reliability judgment of reviews. We also find that not every feature has the same effect in different languages. Additionally, we show that word embedding improves model performance, but significant improvements may require high dimensional word embedding information.

1 Introduction

The increasing number of internet users and widespread participation on social media platforms have led to the creation of various types of content. Reviews are among those content types and their number is continuously growing, as is the number of their readers. Reviews, however, cannot be totally monitored due to their sheer number and the open nature of social media. This impossibility of monitoring them poses an issue, as reviews can be exaggerated or even be written by individuals who have never used the products in question in order to boost the company's sales, rather than by genuine product users. Media consumers may, thus, often find themselves in the situation of being unable to discern between these types of reviews.

Previous studies have explored the characteristics and trained models to distinguish between truthful and deceptive opinion spam, but the results obtained often depended on the dataset used. Different approaches have been employed in developing such datasets. For example, using human judgment to identify deceptive opinions (Li et al., 2011), hiring Amazon Mechanical Turk (AMT) workers to write deceptive reviews (Ott et al., 2011; Harris, 2012), or having experts in the field generate deceptive content (Yoo and Gretzel, 2009) are all approaches that have been adopted. This variety of methods has led to different characterizations of deceptive opinion spam and a debate that is still ongoing. Furthermore, even though a number of features have been used to improve models' performance, most models mainly rely on traditional features related to word usage or to the surface form of the text. Such features are unable to capture the global meaning of the words, which is much more complex and potentially important for distinguishing between truthful and deceptive reviews. Finally, most studies are primarily focused on English-language data, a fact that raises the question of whether the characteristics and features that characterize deceptive opinions in English would do so in typologically different languages.

Given this background, in this paper we tackle the problem of deceptive opinions from a different angle, using different features, and in a typologically different language, like Thai. We try to model directly how social media users may decide whether a product review may be reliable or not. Additionally, we examine the role of lexical content via the incorporation of word embedding, dense vector representations of words that can represent their semantic relationships (Mikolov et al., 2013) into our models. Finally, we conduct our study on Thai, a language that is both typologically very different and relatively understudied when compared to English.

Our study is complementary to previous work

in several aspects. First, we do not rely on generated data with gold labels, but rather we model directly how users reach a judgment of reliability for a given Thai review tweet by building a simple logistic regression model that is intended to mimic their judgments and that is able to achieve reasonable accuracy on this task. The simplicity of logistic regression allows us to assess the roles of different linguistic and paralinguistic features in a typologically different language and check whether the observed patterns resemble those reported when attempting to generate deceptive reviews through methods such as using AMT or involving experts. The addition of word embedding for this type of task, which to our knowledge has not been attempted for related tasks in Thai, allows us to try establishing a role for lexical, semantic, and syntactic information in tasks like deceptive opinion identification.

2 Previous Work

Deceptive opinion spam, a fictitious opinion that has been deliberately written to sound authentic (Ott et al., 2011), has been studied in terms of its characteristics and various approaches aiming at detecting it. Ott et al. (2011) found that humans tend to perform poorly in this task and presented an alternative approach developed on a human-created dataset obtained by employing the services of AMT on a crowdsourcing platform to generate deceptive reviews. This approach is costly, time-consuming, and it still fails to fully reflect real-world scenarios (Ren and Ji, 2019), as the result regarding spatial detail in truthful reviews is different from those obtained on the basis of the Yoo and Gretzel (2009)'s dataset, where the deceptive reviews are from experts in tourism marketing. If the reviewing authors are experienced in that field, providing in-depth details to write a deceptive review is not a challenging task. As a consequence, distinguishing between truthful and deceptive reviews is even more difficult in datasets where deceptive opinions are generated by domain-experts.

A variety of features have been used to detect deceptive opinion spam. Examples of text-based features are the following: Linguistic Inquiry and Word Count (LIWC), which captures both linguistic and psycholinguistic aspects (Ott et al., 2011); Part of Speech tagging, which has been found to have different distributions between truthful and deceptive reviews (Ott et al., 2011); stylometric features, which can reflect the writing style of reviewers even when they try to produce deceptive content (Shojaee et al., 2013); and Bag-of-Words (BoW) representations, which examine the occurrence of words (Ott et al., 2011; Songram et al., 2016). It is clear from the review above that the use of these features primarily focuses on the very surface structure of the text and still lacks the ability to capture specific linguistic and paralinguistic aspects, as well as the lexical, syntactic, and semantic properties of the review at hand.

An additional limitation is that most existing studies have focused on English reviews. For Thai, for instance, we are only aware of a single study conducted to explore deceptive job application messages (Songram et al., 2016). The work in question, just like previous work on English, still relied on examining the occurrence of specific words in the message without capturing the global characteristics of the review, and it did not try to model the relative importance of different features either.

To further our understanding of how to discern reliable from unreliable social media content, we propose a method for detecting the credibility of reviews based on the direct perspective of readers, which can reflect the readers' thoughts and interpretations of reviews in the real world. We also incorporate word embedding to help capture the semantic meaning of the text while continuing to explore various linguistic features to understand the differences in characteristics that influence the credibility of reviews in both Thai and other languages.

3 Approach

3.1 Dataset preparation

Our dataset consists of 239 Thai tweets selected on the basis of their features. First, we selected tweets based on consistency of content, limited to beauty product reviews on Twitter. Second, we also controlled the dataset so that every review must contain the brand name of the product being reviewed. These choices were necessary to ensure that we are modeling differences in reliability rather than content itself and to ensure a similar feature set for all reviews.

The reliability of the reviews was judged by three undergraduate students studying at a Thai university. Each participant was asked to read the review and judge whether it was reliable for them

Feature name	Description
The use of URLs	Whether or not the tweet contains URLs
The use of transliterated brand name	Whether or not the tweet contains transliterated brand name
Number of hashtags	The number of hashtags contained in the tweet
Number of emojis	The number of emojis contained in the tweet
Number of first pronouns	The number of first pronouns contained in the tweet
Number of words	The number of words contained in the tweet
Number of unique words	The number of unique words contained in the tweet
Repetition of characters	Whether or not the tweet contains the repetitive characters
Repetition of exclamation marks	Whether or not the tweet contains exclamation marks

Table 1: Hand-picked features used as predictors of the models

or not. For the gold label, we applied the majority voting rule as described in Li et al. (2011) and Ott et al. (2011). This rule selects the label based on the agreement of two judges to reduce the bias that may arise from human judges. The agreement between judges is only slight according to Fleiss' kappa measure (κ = .04), similar to previous work (Ott et al., 2011),

Of the 239 reviews, 159 were judged as reliable and 80 as unreliable. In order to mitigate the biases in the distribution of "reliable" and "unreliable" reviews, which are known to affect models like Logistic Regression that we use in our research, we conducted experiments by oversampling the unreliable reviews to have the same amount as reliable reviews. Our oversampled dataset consists of 159 reviews for category, i.e., "reliable" and "unreliable" reviews.

3.2 Feature Extraction

To examine features that contribute to the reliability of the reviews, we not only adopted different features that have been identified in previous work but also created new features adapted to the Thai language and to social media-style writing. These include the use of transliterated brand names, the repetition of characters at the end of words, and the number of hashtags. Since Thai is not the origin of many brands, instead of using the percentage of brands mentioned in the reviews used by Yoo and Gretzel (2009), we made a modification and examined whether containing brand names in transliterated form or correctly spelled English form has an effect on the reliability of the reviews. Moreover, in Thai, the repetition of letters at the end of words is often used to emphasize or intensify the meaning of those words. Thus, using this feature can help identify the use of exaggerated expressions, which was found to have a significant effect in Ott et al. (2011)'s work. The use of hashtags is a relatively recent development in the realm of online communication and has been used for various purposes, including acting as a topic-making function by categorizing and organizing content to make it easier to find a post related to a specific topic or trend. Furthermore, hashtags can also serve as a specific type of punctuation to indicate that a tag is metadata, as found in Zappavigna (2015)'s work. All of the features used in our work are summarized in Table 1.

As for comparing model performance, we included one more feature, word embedding, to capture the content of each review. The word embedding can help machines understand the complexity of human language by mapping the meaning of words into a relatively high-dimensional vector space, as shown in Figure 1. The distance of pairs "อ่อนโยน" (gentle) and "ปลอบประโลม" (comfort) is not far from each other due to the similarity of the words; however, the distance of other word pairs "ชุ่มชื้น" (moist) and "แห้ง" (dry) is larger than the former one since they have opposite meanings. We hypothesize that incorporating the semantic context by including word embedding may help our model improve its performance.

The word embedding we used is static word embedding from the Thai National Corpus (TNC) with different dimensionality; we compared models with 50-dimension, 100-dimension, and 200-dimension word embedding. To make use of word embedding, we transformed each word in the reviews into embedding vectors and then averaged all the word vectors in that review to be just one value for each vector dimension. Therefore, one review will have 50, 100, or 200 embedding values, depending on the model's architecture.

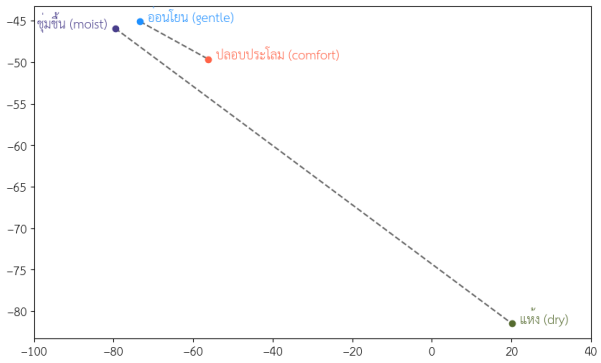

Figure 1: The example of the pairwise word distances in a two-dimensional embedding space reduced using t-Distributed Stochastic Neighbor Embedding (t-SNE)

	Estimate	SE	p-value
(Intercept)	-0.63	0.59	0.29
URLs	-2.07	0.50	<0.005
transliterated	-0.07	0.26	0.78
numHashtag	-0.64	0.16	<0.005
numEmoji	0.11	0.12	0.36
numFirstPron	0.75	0.23	<0.05
numWord	0.04	0.03	0.19
numUniqueWord	-0.02	0.04	0.56
Repetition	0.01	0.26	0.96
ExclamationMark	-1.27	0.32	<0.005

Table 2: Coefficients, standard error and p-value of the logistic regression model with features as independent variables

3.3 Experiment Setup

Our data was analyzed by fitting logistic regression which consist of two types of models. The first model had nine hand-picked features as predictors: `Reliability ~ Features`, the other models were similar to the first one, but we also added word embedding as predictors: `Reliability ~ Features + Word Embedding`.

Despite fitting the models with three different amounts of dimension to evaluate the performance of classifying the reviews' reliability, only 50-dimension embedding was implemented to identify the coefficients of our hand-picked features. This choice was made to mitigate collinearity between word embedding and other predictors as well as convergence issues with the highest number of dimensions.

Additionally, to ensure the reliability of our models, we employed 5-fold cross-validation to evaluate their performance and identify the most reliable results.

4 Results

4.1 Model with only linguistic features as predictors

A logistic regression analysis was performed to assess the effects of the linguistic and paralinguistic features on the likelihood of the reviews being deceptive. The logistic regression model was an improvement over an intercept-only model, $\chi^2(308) = 67.2$, $p < .001$. It was found that four features were significant, including the use of URLs, the number of hashtags, the number of first-person pronouns, and the use of exclamation marks, as reported in Table 2.

The result indicated (as shown in Figure 2) that the use of URLs was associated with an increased likelihood of the reviews' deceptiveness; the reviews with more hashtags had a higher probability of being unreliable than the ones with fewer hashtags; the reviews with more first-person pronouns were more likely to be reliable; and the use of exclamation marks negatively affected the reliability of the reviews. Additionally, the use of transliterated forms of brand names, the number of emoji, the length of words, the repetition of the characters, and the number of unique words were not associated with changes in the likelihood of the reviews' reliability.

4.2 Model with features and word embedding as predictors

For the model including 50-dimension word embedding as predictors, we also performed a logistic regression analysis to assess both the effects of the linguistic and paralinguistic features and the effects of word embedding on the likelihood of the reviews being deceptive. The logistic regression model was an improvement over an intercept-only model, $\chi^2(258) = 188$, $p < .001$. It was found that three features were significant, including the number of hashtags, the number of first-person pronouns, and the use of exclamation marks, as reported in Table 3.

The finding revealed that including word embedding as a predictor in the model produced a difference in the results. The use of URLs was no longer associated with changes in the likelihood of the reviews' reliability.

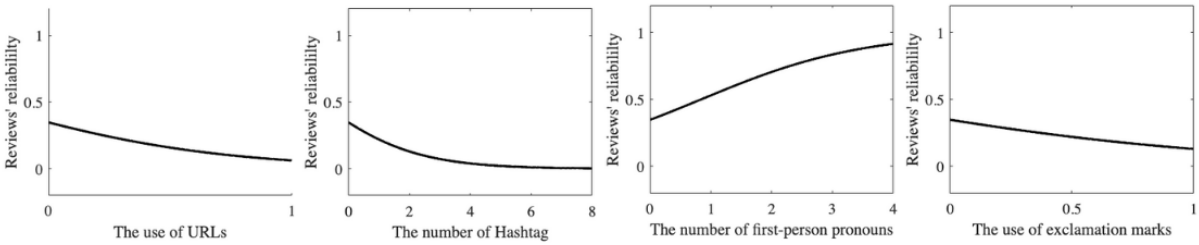

Figure 2: The direction of the relationship between reviews' reliability and features that are significant (p<0.05)

	Estimate	SE	p-value
(Intercept)	-9.74	7.18	0.17
URLs	-1.38	0.91	0.13
transliterated	-0.10	0.41	0.80
numHashtag	-1.44	0.35	<0.005
numEmoji	-0.02	0.18	0.93
numFirstPron	1.42	0.37	<0.05
numWord	0.09	0.05	0.07
numUniqueWord	-8.16	0.07	0.10
Repetition	-0.42	0.41	0.30
ExclamationMark	-2.83	0.63	<0.005

Table 3: Coefficients, standard error and p-value of the logistic regression model with features and word embedding as independent variables

4.3 Model Comparison

The performance of four logistic regression models was evaluated using 5-fold cross-validation. Model 4, with features and 200-dimension word embedding as predictors, had the best scores in every metric: precision, recall, and F1.

We also attempted the classification with other models that may be less sensitive to assumption on data distribution, such as Support Vector Machines, and obtained similar result. We, thus, omit their discussion in view of space limitations.

5 Discussion

After comparing the models on precision, recall, and F1 scores, the findings showed that the model with both features based on domain-specific knowledge and 200-dimension word embedding as predictors had the best performance. This could be due to the additional information contained in word embedding, that is lexical, syntactic, and semantic content.

However, the addition of word embedding did not improve the performance of the model as much as we expected, given the increase in the number of model parameters. This raises the question of whether our choice of linguistic and paralinguistic features may have already served as a way to encode reliability in most aspects, except those dealing with more fine-grained semantic characteristics, which could be covered by word embedding. In other words, the highest performance score from Model 4 with 200-dimension word embedding could be due to more fine-grained semantic content that our hand-picked features did not cover and was supplied by the word embedding instead.

Furthermore, the results demonstrated that some features were able to significantly affect the reliability of the reviews both with or without word embedding, such as the number of hashtags, the number of first-person pronouns, and the use of exclamation marks, but some features were different across the models we experimented with. In particular, this is the case for the use of URLs.

5.1 Number of hashtags

The study of Wadhwa et al. (2017) found that tweets containing a hashtag experience a threefold higher rate of engagement compared to tweets without any hashtags. Due to the fact that hashtags draw more attention from Twitter users, reviewers tend to include hashtags in their tweets. Even though the presence of hashtags is beneficial, tweets with too many hashtags can be seen as spam. As a result, our findings indicate that there is a negative relationship between the number of hashtags present in the reviews and their level of reliability.

5.2 Number of first-person pronouns

Our findings revealed that the reviews with more first-person pronouns were less likely to be unreliable. At first sight, our finding may appear to contrast with the studies by Yoo and Gretzel (2009) and Harris (2012) suggesting that the deceptive reviews have a higher likelihood of containing self-references. In this respect, it is interesting to note that previous studies about self-

	Precision	Recall	F1
Model 1: Features	0.70	0.69	0.68
Model 2: Features + WE50D	0.69	0.69	0.69
Model 3: Features + WE100D	0.69	0.68	0.68
Model 4: Features + WE200D	**0.74**	**0.73**	**0.73**

Table 4: Model Performance Evaluation to classify reliability of the reviews from 5-folds cross-validation of the logistic regression models

references in lies revealed that self-references are used less often in lies because liars tend to disassociate themselves from their lies (Newman et al., 2003). This could offer a perspective unifying previous results as well as ours. Since self-references are avoided when lying, these could naturally be associated with reliable reviews, as in our work, but they may also be exploited by users writing deceptive reviews, as pointed out in previous work.

5.3 The use of URLs

The study of Cholprasertsuk et al. (2020) found that some social media influencers (SMIs) are partnering up with businesses to create new forms of advertisements on social media to boost the company's sales volumes. Therefore, affiliate marketing, one of the new ways to advertise on the internet, is widespread among social media influencers (SMIs) on Twitter. Affiliate marketing happens when reviewers get passive income from links clicked and purchases made through their tweets.

In this study, we used the use of URLs as a paralinguistic feature to identify the reliability of the reviews, as we expected that the reviews with links or URLs would be deceptive, which was true in the first model without word embedding. However, in the model with word embedding, the use of URLs was no longer associated with the reliability of the reviews. This raises the question of whether the expected effect is indeed real.

We conducted a correlation analysis between two variables: the use of URLs and word embedding (50-dimension). Our findings revealed that the use of URLs in reviews and some dimensions of word embedding in our data were moderately correlated. For instance, as shown in Figure 3, the use of URLs and the 15th dimension of word embedding were negatively correlated, $r(316) = -0.43$, $p < .001$.

Due to correlation, the change in the result across the models can be attributed to including word embedding in the presence of correlation between words and the use of links. In other words, the model relied more on word embedding and disregarded the effect of URLs, as their information was already incorporated into another independent variable.

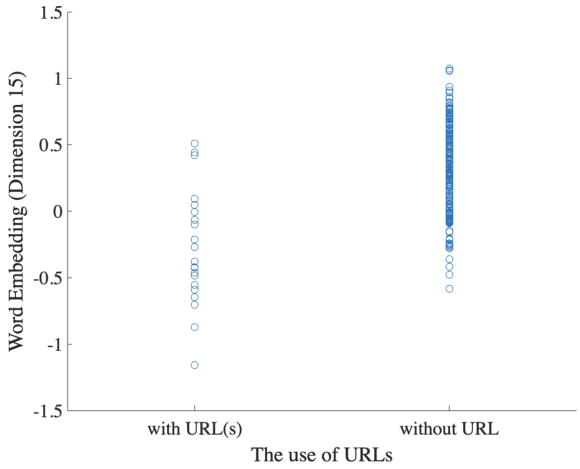

Figure 3: The scatterplot of the use of URLs and the 15th dimension of word embedding

5.4 The use of exclamation marks

It was found that the use of exclamation marks was associated with a lower reliability of the reviews in both models, with or without word embedding. Since we use exclamation marks to emphasize something to be more intense, in this context it could also be used to exaggerate things as well. According to the study of Ott et al. (2011), more exaggerated reviews are less reliable, a fact that aligns with our findings.

6 Conclusion

In this study, we took a different approach to detecting reliable and unreliable Thai Twitter reviews, with a specific focus on the domain of beauty products. Our work offers social media-related findings by exploring new linguistic and paralinguistic features, such as the use of URLs, the presence of emojis, and the number of hashtags,

in the analysis of review reliability. We also discussed how the role of these features may be interpreted in the wider social context of social media usage. Additionally, we incorporated the semantic context of the messages by using word embedding to detect deceptive reviews as well. We found that there exists a clear association between specific linguistic and paralinguistic features and humans' judgments of the reliability of text encountered on the internet. It also appears that the implications of linguistic features are not universal. For example, in this study, the number of first-person pronouns is associated with more reliable tweets in contrast with more deceptive opinions reported for English (Ott et al., 2011). The difference, as we have suggested, could be partly associated with a difference in tasks for which the models have been trained, yet such a finding calls for caution when interpreting the generality of one's findings. The importance of different linguistic features may not be the same in different languages, and it may be modulated by evolving practices and usage in a community of users, as is the case for the number of hashtags. Finally, the results of the models we implemented suggest that word embedding may help supplement the information available to models, yet higher dimensionality and injecting more fine-grained linguistic information may be necessary for more significant improvements in performance.

7 Limitations and Future Work

While this study provides some insights into the importance of features and semantics to predict the reliability of reviews, some limitations need to be acknowledged. First, the domain of reviews is limited, and future work should focus on examining whether our findings extend to tweets with very different contents. Second, our reviews were judged by humans, who may have some bias toward different products and only partially agree with each other. These facts make modeling their reliability ratings challenging. Third, there are limitations related to the models implemented in this study. Logistic regression was chosen to be able to easily and directly ascertain feature weights, yet it has limits when it comes to the models' ability to make use of word embedding to capture complex information contained in the text, as averaging or other strategies are needed to deal with different review lengths. Neural models would probably ensure a better performance and the ability to make use of word embedding information, at the cost of uncertainty regarding the role of different linguistic features. Last, there are a few differences between our work and previous studies which can be due to many reasons such as the statistical model chosen and the domain of the dataset. Most previous work experimented with Naive Bayes or Support Vector Machine Model (SVM) while we chose to perform Logistic Regression. In addition, the domain of our data, beauty reviews, is quite different compared to other studies.

Future work should try to extend the approach presented to novel domains, rely on a larger number of annotators, and complement logistic regression with models that can overcome the limitations previously discussed. We also expect that work on different languages and in different communities of users may showcase the language- and context-dependent nature of the effect of linguistic and paralinguistic features on human reliability judgments.

Acknowledgements

We thank fellow classmates, instructors, and teacher assistants from the Quantitative Methods in the study of Language course taught at Chulalongkorn University in the Spring of 2023 for providing valuable comments and advice on earlier versions of this work.

References

Anchalee Cholprasertsuk, Chayakorn Lawanwisut, and Sirinart Thongrin. 2020. Social media influencers and thai tourism industry: Tourists' behavior, travel motivation, and influencing factors. *Journal of Liberal Arts, Thammasat University*, 20(2):234–263.

C. G. Harris. 2012. Detecting deceptive opinion spam using human computation. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

F. H. Li, M. Huang, Y. Yang, and X. Zhu. 2011. Learning to identify review spam. In *Twenty-second International Joint Conference on Artificial Intelligence*.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Matthew Newman, James Pennebaker, Diane Berry, and Jane Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality Social Psychology Bulletin*, 29:665–675.

M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

Y. Ren and D. Ji. 2019. Learning to detect deceptive opinion spam: A survey. *IEEE Access*, 7:42934–42945.

S. Shojaee, M. A. A. Murad, A. B. Azman, N. M. Sharef, and S. Nadali. 2013. Detecting deceptive reviews using lexical and syntactic features. In *2013 13th International Conference on Intelligent Systems Design and Applications*, pages 53–58.

P. Songram, A. Choompol, P. Thipsanthia, and V. Boonjing. 2016. Detecting thai messages leading to deception on facebook. In *Integrated Uncertainty in Knowledge Modelling and Decision Making: 5th International Symposium, IUKM 2016, Da Nang, Vietnam, November 30-December 2, 2016, Proceedings 5*, pages 293–304. Springer International Publishing.

V. Wadhwa, E. Latimer, K. Chatterjee, J. McCarty, and R. T. Fitzgerald. 2017. Maximizing the tweet engagement rate in academia: Analysis of the ajnr twitter feed. *AJNR. American Journal of Neuroradiology*, 38(10):1866–1868.

K. H. Yoo and U. Gretzel. 2009. Comparison of deceptive and truthful travel reviews. In *Information and Communication Technologies in Tourism*, pages 37–47.

Michele Zappavigna. 2015. Searchable talk: The linguistic functions of hashtags. *Social Semiotics*, 25(3):274–291.