International Journal of Information Management Data Insights 2 (2022) 100123

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei

A comparative analysis of followers' engagements on bilingual tweets using regression-text mining approach. A case of Tanzanian-based airlines

Boniphace Kutela[a,*], Raynard Tom Magehema[b], Neema Langa[c], Felistus Steven[d], Rafael John Mwekh'iga[b]

[a] Roadway Safety Program, Texas A&M Transportation Institute, 1111 RELLIS Parkway, Bryan, TX 77807, United States
[b] Department of Civil Engineering, Ardhi University, P.O. Box 35176, Dar es salaam, Tanzania
[c] Department of Sociology/African American Studies, University of Houston, 3551 Cullen Boulevard, Houston, TX 77204, United States
[d] Archbishop Mihayo University College of Tabora, P.O. Box 801, Tabora, Tanzania

ARTICLE INFO

Keywords:
Multiple languages
Social media engagement
Airline marketing
Text-mining

ABSTRACT

Business entities utilize multiple languages on social media for marketing, promotions, and communication. The impact of utilizing one language over the other on engagements has not been well explored. This study applied a regression-text mining approach to over 3000 Tanzanian-based airlines' tweets posted between 2018 and 2020 to explore the influence of English and Swahili languages in communication and marketing. By defining engagement as the retweets and favorites, the study found that English tweets had relatively higher engagements than Swahili tweets. However, Swahili tweets with photos and videos were likely to have more engagements than English tweets. Conversely, hashtags attracted higher engagements for English tweets. Time of the day revealed mixed findings. Further, several key patterns were observed when favorites and retweets were considered separately. The practical applications of the study were also discussed. It is expected that the study findings will benefit bilingual business entities on a global scale.

1. Introduction

Social media platforms enable quick electronic communication of substances, including personal information, documents, videos, and photos. The use of social media, such as Facebook, Instagram, and Twitter, has increased rapidly worldwide (Ortiz-Ospina, 2019). Approximately one-third of the world's population currently uses social media (Zahoor and Qureshi, 2017). The increased number of social media users has triggered businesses to take advantage of these platforms for marketing and interaction with customers, thus making social media an effective marketing tool (Zahoor and Qureshi, 2017; Ashley and Tuten, 2015; Oliverio, 2018; Hanaysha, 2022). Businesses use social media to learn and map their customers' behaviors through online information queries, reviews, and promotions (Hanaysha, 2022; Zeng and Gerritsen, 2014). The benefits of social media marketing for businesses include increasing sales, generating traffic to online platforms, increasing brand awareness, improving brand image, lowering marketing costs, and increasing user interaction on platforms by incentivizing users to share content (Felix et al., 2017).

Airlines across the globe have been active in social media, such as Twitter, to communicate and promote their deals to their followers and the public in general. As people in various nations use more than one language for communication, airlines' Twitter handles also tend to interact with their followers using multiple languages. For instance, airlines' Twitter handles in the United States are likely to use English and Spanish, Canada uses English and French, Rwanda uses Kinyarwanda and French, and Tanzania uses Kiswahili and English, among other nationals. The use of a specific language intends to target a group of customers who can easily understand it to maximize engagement. However, it is not known whether the use of one language over the other tends to create more engagement on social media platforms.

Therefore, this study utilizes Twitter data from Tanzania to explore the influence of languages on customer engagement. The key questions to be answered in this study are; (1). Does the use of local and foreign languages on Twitter as a social media platform for communication or advertisement lead to the same number of followers' engagements? (2). Are followers' engagements different if the two languages are used together with other features such as hashtags and mentions? (3). Does the content of the two languages differ significantly?

The findings from this study would provide insights that can be applied to business entities to improve interactions with customers by focusing on the specific language of choice. Specifically, this study

* Corresponding author at: Roadway Safety Program, Texas A&M Transportation Institute, United States.
E-mail address: b-kutela@tti.tamu.edu (B. Kutela).

https://doi.org/10.1016/j.jjimei.2022.100123
Received 26 February 2022; Received in revised form 22 September 2022; Accepted 22 September 2022

presents the key factors associated with the increased engagements in social media that business entities can utilize to improve their interactions with customers. Further, the content analysis presented in this study provide a better understanding of the specific language's phrases that when used in social media improve customers engagement.

The remainder of the manuscript is organized as follows; the next section presents the literature review followed by methodology adopted and data description. The results and the discussion are then presented, followed by a conclusion as well as the future studies section.

2. Literature review

This section presents the literature review. It covers the social media business in cooperation, choice of language in social media, and methodology and techniques used in previous studies.

2.1. Social media use by business entities

Approximately 93% of businesses worldwide have incorporated social media platforms to interact with customers (Alalwan et al., 2017). For example, Facebook and Twitter provide a platform for linking business advertisements within customer interfaces (Oliverio, 2018). The ability of social media to use visual, verbal, and textual content interactions simultaneously has allowed businesses to engage with customers globally, despite cultural and language differences (Whitelock et al., 2013). Social media platforms have developed tools that enable real-time multilingual engagements; for example, over 70 translations have been created by translators and volunteers on Facebook (Singh et al., 2012). However, increased preference for the use of local languages has posed a challenge to social media platforms' efforts to address multilingual engagements. Almost 40% of global social media users prefer content in languages other than English, but only 23% translate their social media content (Singh et al., 2012). Multilingualism also segregates information and communication domains based on language, limiting the potential for intercultural dialogue and transnational collaboration (Eleta and Golbeck, 2014).

2.2. Choice of language on social media

Often social media platforms are based on the English language, and businesses prefer to use English to engage with most customers. However, customers are engaging with companies in diverse languages. Scholars have presented various theoretical and practical implications of multilingual social media marketing and engagements. Sass et al. (2020) contend that consumer perceptions are independent of language and found product perceptions were similar among English, French, Spanish, and Portuguese Twitter users. The chi-square test identified significant differences in the frequency of tweets in each category between languages. The correspondence analysis multivariate technique was used to measure and visualize the similarity between selected tweets (Sass et al., 2020). A multilingual online environment enables users unfamiliar with international languages to access content, while the ability to switch between languages and translations is required to overcome language barriers (Eleta and Golbeck, 2014). However, an experiment on native English-speaking Facebook users revealed that the language used in posts significantly affects engagement, interaction, and reaction (Lim and Fussell, 2017). Using a two-by-two factorial within-subject design, Lim and Fussell (2017) discovered that English-speaking Facebook users engaged more with English-language posts than foreign-language posts regarding favorites, comments, and time spent.

2.3. Methodology and techniques used in previous studies

A large amount of consumer reaction data in the form of tweets, favorites, retweets, and comments stored in the Twitter database, has piqued the interest of several researchers to conduct what is popularly referred to as Twitter mining. Jung & Jeong developed a machine learning model using Twitter-generated social media features to predict start-up firms' social media engagement levels, noting that likes, retweets, and comments influenced firms' engagement levels (Jung and Jeong, 2020). Mazumdar & Thakker (2020) used exploratory visual analysis to map the most frequently discussed topics related to citizen science in Twitter microblogging communities. A few studies have examined the influence of multilingualism on social media engagement. However, the approaches used in these studies do not uncover the influence of various tweet contents on consumer engagement. For instance, Aswani et al. (2017) used text mining to explore the content virality of Facebook posts. They found that a post going viral depends on factors such as motivating consumers to like/comment/share a post. However, Aswani et al. (2017) also point out the uncertainty of these posts getting the desired popularity, and further studies are needed.

To this end, various studies have explored followers' engagements on social media platforms. Various methodologies have been applied to explore engagements on social media platforms. However, machine learning-based methods alone tend to be a black box, meaning the reader does not understand what happened behind the scenes. On the other hand, regression-based methods tend not to explain the causation but associations between engagements and predictors. Thus, this study uses the hybrid (regression-text) model to understand the customer engagements for multilingual tweets posted by the airlines' Twitter handles in Tanzania.

3. Methodology

This section presents data collection and analysis approaches applied in this study. This study utilized tweets from Tanzanian-based airlines to explore the influence of one language over the other on followers' engagements. Two approaches, Negative Binomial (NB) regression and text network analysis (TNA), were used for data analysis. First, since engagement is measured by the number of favorites and retweets, the NB regression was used to estimate the influence of various predictors on followers' engagement. Secondly, the TNA was used to understand the key tweet contents that drive the engagements. The following section provides in detail the data collection procedure and theoretical background of NB and TNA.

3.1. Data collection

The study uses three years (2018-2020) of publicly available tweets from two Tanzanian-based airlines' handles (Air Tanzania and Precision Air) to explore the influence of language used for communication on followers' engagements. The followers' engagements are measured by the number of favorites (likes) and retweets. During that time, the two accounts had a total of 3276 tweets, 788 replies, and 302 retweets. Amongst, 1887 tweets, 255 replies, and 139 retweets were communicated in English, while 1388 tweets, 533 replies, and 163 retweets were communicated in Swahili. The tweets were retweeted and liked 97289 and 17565 times, respectively. This study focused only on the tweets from these two accounts since retweets and replies might not be intended for communication purposes. Various tweet characteristics such as tweet type, multimedia type in the tweet, number of mentions or tags in a tweet, and hashtags in the tweet were explored. The number of words in the tweet, language used, year of the tweet, day of the week, and time of the day were also explored. Table 1 presents the descriptive statistics of the explanatory variables. For the multimedia criteria, English tweets with photos had the most significant proportion of favorites (82.58%) and retweets (79.59%). Furthermore, tweets with no mentions constitute the largest proportion for both English and Swahili tweets, irrespective of the engagement type.

On the other hand, tweets with three or more mentions had the lowest engagement (favorites and retweets) among all tweets. For example,

Table 1
Descriptive statistics of the explanatory variables.

	Total engagements		Favorites only		Retweets Only	
	English	Swahili	English	Swahili	English	Swahili
Multimedia type						
No multimedia	3.19%	7.41%	3.05%	7.23%	3.96%	8.36%
Photo	82.13%	80.56%	82.58%	81.09%	79.59%	77.75%
Video	14.68%	12.03%	14.37%	11.68%	16.45%	13.90%
Number of hashtags						
No hashtag	6.70%	24.77%	6.44%	24.74%	8.18%	24.97%
One hashtag	6.91%	8.62%	6.68%	8.58%	8.23%	8.82%
Two hashtags	17.56%	16.76%	17.84%	16.95%	15.97%	15.72%
Three hashtags	34.98%	25.85%	35.36%	25.99%	32.79%	25.08%
Four hashtags	16.16%	12.67%	16.26%	12.37%	15.58%	14.22%
Five or more	17.69%	11.33%	17.42%	11.36%	19.24%	11.18%
Number of mentions						
No mention	94.21%	94.46%	94.25%	94.59%	93.99%	93.74%
One mention	4.88%	3.71%	4.87%	3.63%	4.92%	4.11%
Two mentions	0.72%	1.35%	0.70%	1.34%	0.84%	1.40%
Three or more	0.19%	0.48%	0.18%	0.43%	0.25%	0.75%
Time of the day						
Early morning (7–9am)	54.82%	36.99%	55.56%	37.29%	50.57%	35.40%
Late morning (9am–12pm)	13.55%	17.77%	13.24%	17.54%	15.35%	18.97%
Afternoon (12–3pm)	7.43%	15.46%	7.24%	15.33%	8.52%	16.13%
Evening (3–6pm)	19.28%	21.59%	19.06%	21.65%	20.56%	21.26%
Nighttime (6pm–7am)	4.92%	8.20%	4.90%	8.19%	5.01%	8.24%
Day of the week						
Weekday	84.72%	80.17%	84.73%	80.06%	84.66%	80.71%
Weekend	15.28%	19.83%	15.27%	19.94%	15.34%	19.29%
Year of the tweet						
2018	18.06%	38.06%	16.95%	37.01%	24.36%	43.63%
2019	48.75%	39.73%	48.88%	40.06%	48.00%	37.93%
2020	33.19%	22.22%	34.17%	22.93%	27.64%	18.44%

English tweets had only 0.25% of retweets and 0.18% of favorites compared to Swahili tweets, which had only 0.75% of retweets and 0.43% of favorites. Moreover, English tweets during weekdays had 84.73% of favorites, while Swahili had 80.06 % of total favorites. Additionally, tweets in the nighttime have the lowest share across all engagements, followed by tweets in the afternoon, irrespective of the language. English tweets in 2019 had a (48.88%) proportion of favorites compared to other years. The same is noticed for Swahili tweets' 40.06% in the same year compared to other years; however, in 2018, Swahili tweets had the largest share of retweets (43.63%).

3.2. Negative binomial regression

The study develops a model to estimate followers' engagement y_i given the characteristics of the tweet X_i. The response variable (number of favorites and retweets) is non-continuous and cannot take negative values (Cox, 1983; Kutela and Teng, 2019; Gardner et al., 1995). Thus, the family of count models is the most appropriate for data analysis. If μ is the mean number of engagements on a tweet such that it is assumed to follow Poisson distribution, then,

$$f(y_i|X_i) = \frac{e^{\mu i}{\mu_i}^{yi}}{y_i!} \tag{1}$$

However, since some factors cannot be measured, then the model translates to overdispersion in which the variance is greater than the mean (Cox, 1983; Kutela and Teng, 2019). The Poisson distribution is automatically violated; hence negative binomial model should be used. This model is suitable for overdispersion where parameter α, is introduced and the model becomes (Booth et al., 2003):

$$Pr(Y_i = y_i; \alpha, \mu_i) = \frac{\Gamma(y_i + \alpha)}{\Gamma(\alpha) y_i!}\left(\frac{\alpha}{\mu_i + \alpha}\right)^{\alpha}\left(\frac{\mu_i}{\mu_i + \alpha}\right)^{y_i} \tag{2}$$

It follows that the expected mean and variance are computed as μ_i and $\mu_i + {\mu_i}^2/_{\alpha}$ and a log-linear relationship between engagement on a tweet (favorites, retweets, or both) y_i and characteristics of the tweet X_i exists. The log-linear relationship makes it possible to use multiple repressors, therefore,

$$ln(E(y_i|x_i) = \ln\left(\mu_i\right) = X_i\beta \tag{3}$$

Where, X_i denotes the repressors, y_i is the tweet engagement, and β is the matrix of regression coefficients.

Log-likelihood maximization is used in estimating the coefficients and dispersion parameters, and it is then added to get the likelihood of all tweet languages.

3.3. Text network analysis

Text network analysis (Yoon and Park, 2004; Kushwaha et al., 2021; Benoit et al., 2018) is a technique that reveals a hidden relationship between textual data using nodes and edges (Yoon and Park, 2004; Kushwaha et al., 2021; Kutela et al., 2022). It has extensively been used in linguistics and literature reviews to understand the hidden patterns in textual data (Kushwaha et al., 2021; Saheb et al., 2021; Hunter, 2014; Jiang et al., 2020; Kumar et al., 2021). It has also proved efficient in analyzing big data and literature reviews (Kushwaha et al., 2021). In simple terms, text networks constitute the nodes, which are the keywords, and edges, which show the co-occurrence relationship between the keywords in a sentence. Such edges increase in thickness as the number of the same paired keywords increases. Fig. 1 shows the typical text network's major components, which include nodes, edges, and community. A community is a cluster of keywords with similar themes (Kim and Jang, 2018; Kutela et al., 2022).

The text network analysis involves four major steps, which are (i) text normalization (removal of stop words and changing capital to lower case letters), (ii) creating a corpus of words, (iii) generation of nodes and links, and (iv) development of quantitative indexes for in-depth analysis (Yoon and Park, 2004; Kim and Jang, 2018; Paranyushkin, 2011; Kutela et al., 2021, Kutela et al., 2021). It should be noted that stemming and lemmatization, which are common pre-processing approaches in text mining, were not performed in this study to maintain the origi-

Table 2
Negative Binomial regression results for retweets and favorites.

	Overall			English tweets			Swahili tweets		
	Coef	Exp (coef)	P-value	Coef	Exp (coef)	P-value	Coef	Exp (coef)	P-value
Language									
Swahili tweets									
English tweets	0.051	1.05	0.261						
Multimedia type									
No multimedia									
Photo	1.184	3.27	<0.001	0.718	2.05	<0.001	1.147	3.15	<0.001
Video	1.855	6.39	<0.001	1.438	4.21	<0.001	1.944	6.99	<0.001
Number of hashtags									
No hashtag									
One hashtag	−0.100	0.90	0.204	0.179	1.20	0.103	−0.293	0.75	0.010
Two hashtags	0.103	1.11	0.173	0.323	1.38	0.002	−0.339	0.71	0.002
Three hashtags	0.465	1.59	<0.001	0.750	2.12	<0.001	0.135	1.14	0.194
Four hashtags	0.992	2.70	<0.001	1.115	3.05	<0.001	0.940	2.56	<0.001
Five or more hashtags	1.106	3.02	<0.001	1.330	3.78	<0.001	1.051	2.86	<0.001
Number of mentions									
No mention									
One mention	−1.061	0.35	<0.001	−0.995	0.37	<0.001	−1.813	0.16	<0.001
Two mentions	−0.910	0.40	<0.001	−1.481	0.23	<0.001	−1.128	0.32	<0.001
Three or more mentions	−0.744	0.48	<0.001	−0.812	0.44	0.001	−1.236	0.29	<0.001
Time of the day									
Early morning (7-9am)									
Late morning (9am-12pm)	−0.001	1.00	0.990	0.333	1.40	0.001	−0.006	0.99	0.954
Afternoon (12–3pm)	−0.217	0.80	0.002	−0.016	0.98	0.867	−0.096	0.91	0.342
Evening (3–6pm)	−0.409	0.66	<0.001	−0.443	0.64	<0.001	−0.189	0.83	0.028
Nighttime (6pm–7am)	−0.085	0.92	0.386	0.284	1.33	0.058	−0.104	0.90	0.429
Day of the week									
Weekday									
Weekend	−0.168	0.85	0.002	−0.296	0.74	0.000	0.038	1.04	0.650
Tweet year									
2018									
2019	0.215	1.24	<0.001	0.353	1.42	<0.001	−0.073	0.93	0.378
2020	0.087	1.09	0.150	0.317	1.37	<0.001	−0.435	0.65	<0.001
Ln (Number of words	0.066	1.07	0.054	−0.141	0.87	0.008	0.105	1.11	0.021
Intercept	1.769	5.87	<0.001	2.477	11.90	<0.001	2.100	8.17	<0.001

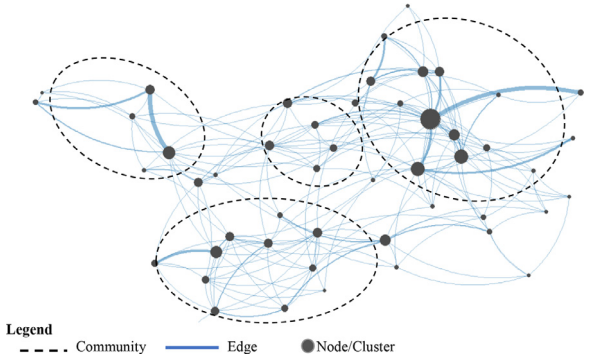

Fig. 1. A conceptual skeleton of the text network (Kutela et al., 2021).

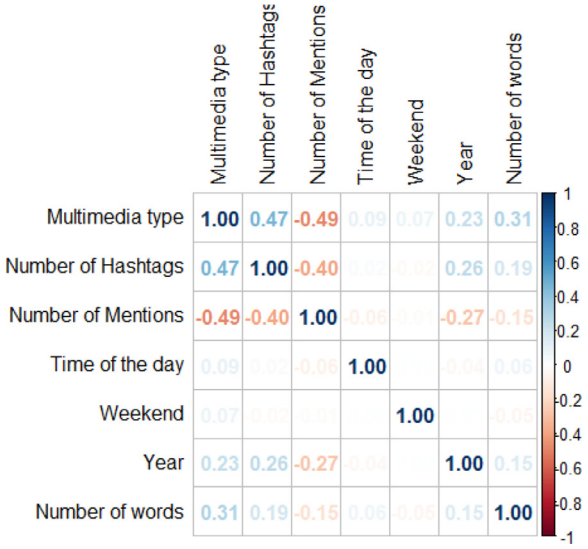

Fig. 2. Variables' correlation coefficients.

nality of the text. The in-depth analysis focused on the network topology and key clusters and communities extracted from the network. In this study, analysis was performed using *quanteda* and *igraph* packages (Benoit et al., 2018; Csárdi, 2020) for creating the text network and extraction of the network parameters. Only the top 40 keywords were considered for interpretation.

4. Results and discussions

The results constitute two parts, the negative binomial regression results and text network results. The NB intended to explore the influence of individual variables on total engagement and individual engagement type. On the other hand, the text network intended to explore the key topics for different factors associated with high engagements.

Before NB regression development, correlation analysis was performed to avoid multicollinearity. Fig. 2 presents the correlation results of the variables in this study. It can be observed that all variables are not highly correlated, and the largest correlation coefficient (−0.49) is between the number of mentions and multimedia type. Therefore, all variables are included in the model.

4.1. Negative Binomial results and discussion

Tables 2–4 show the negative binomial regression analysis results for total engagement (retweets and favorites combined), favorites, and

Table 3
Negative Binomial regression results for favorites.

	Overall			English tweets			Swahili tweets		
Variable	Coef	Exp (coef)	P-value	Coef	Exp (coef)	P-value	Coef	Exp (coef)	P-value
Language									
Swahili tweets									
English tweets	0.046	1.05	0.306						
Multimedia type									
No multimedia									
Photo	1.206	3.34	<0.001	0.899	2.46	<0.001	1.169	3.22	<0.001
Video	1.851	6.36	<0.001	1.592	4.91	<0.001	1.937	6.94	<0.001
Number of hashtags									
No hashtag									
One hashtag	−0.104	0.90	0.186	0.198	1.22	0.069	−0.309	0.73	0.006
Two hashtags	0.110	1.12	0.140	0.351	1.42	0.001	−0.331	0.72	0.002
Three hashtags	0.467	1.60	<0.001	0.751	2.12	<0.001	0.131	1.14	0.203
Four hashtags	0.981	2.67	<0.001	1.109	3.03	<0.001	0.915	2.50	<0.001
Five or more hashtags	1.098	3.00	<0.001	1.318	3.74	<0.001	1.036	2.82	<0.001
Number of mentions									
No mention									
One mention	−1.041	0.35	<0.001	−0.850	0.43	<0.001	−1.783	0.17	<0.001
Two mentions	−0.874	0.42	<0.001	−1.312	0.27	<0.001	−1.085	0.34	<0.001
Three or more mentions	−0.796	0.45	<0.001	−0.655	0.52	0.007	−1.318	0.27	<0.001
Time of the day									
Early morning (7-9am)									
Late morning (9am-12pm)	−0.006	0.79	0.926	0.267	0.92	0.006	−0.004	0.89	0.970
Afternoon (12–3pm)	−0.230	0.67	0.001	−0.084	0.63	0.375	−0.112	0.83	0.259
Evening (3–6pm)	−0.404	0.94	<0.001	−0.460	1.33	<0.001	−0.183	0.90	0.031
Nighttime (6pm–7am)	−0.064	1.00	0.511	0.287	1.00	0.052	−0.100	1.00	0.441
Day of the week									
Weekday									
Weekend	−0.159	0.85	0.003	−0.342	0.71	<0.001	0.052	1.05	0.530
Tweet year									
2018									
2019	0.261	1.30	<0.001	0.447	1.56	<0.001	−0.027	0.97	0.742
2020	0.148	1.16	0.014	0.444	1.56	<0.001	−0.380	0.68	<0.001
Number of words	0.046	1.05	0.178	−0.142	0.87	0.007	0.096	1.10	0.034
Intercept	1.604	4.97	<0.001	2.057	7.82	<0.001	1.906	6.73	<0.001

retweets, respectively. The explanatory variables include language type, multimedia type, number of hashtags, number of mentions, time of the day, day of the week, year of a tweet, and number of words per tweet. The interpretation is based on the expected value, which is the exponential of the coefficients. Further, the 95% confidence level of statistical significance of the variables is considered.

According to the results in Tables 2–4, English tweets are likely to have more engagement (favorites and retweets) than Swahili tweets. However, the variable is not statistically significant at a 95% confidence interval. In other words, the language used in a tweet has no statistically significant impact on followers/public engagements.

Although the language by itself has a statistically insignificant different impact on the engagements, tweet contents have shown a great difference when they are used in two languages. Results show that the type of multimedia is positively associated with engagements. The association is statistically significant at a 95% confidence level, irrespective of the type of engagement and language used. However, the magnitude of the impact varies per multimedia type and language used. The overall assessment results in Table 2 show that tweets with video are about six times more likely to generate engagements than tweets without multimedia. This is likely since videos may contain more information/gestures than plain words (Anand et al., 2021). This is consistent with findings from Ahn et al. (2021), who analyzed public engagement on Twitter using topic modeling.

Further, tweets with pictures are about three times more likely to generate engagement than tweets without multimedia. This means that pictures are less likely to generate engagement than videos. This observation agrees with Csárdi (2020), who reported that well-constructed videos trigger more accurate emotional reactions than pictures. Again, this may be because of the information associated with the media type, for instance, certain gestures from the videos. The results also show that Swahili tweets with videos generate more engagement about (average of seven times) than English tweets (average of four times). The results are consistent for favorite-only analysis (Table 3) and retweets-only analysis (Table 4).

The number of hashtags has revealed mixed outcomes. Swahili tweets with a low number of hashtags (one and two hashtags) are negatively associated with engagements, irrespective of engagement type. This is contrary to Anand et al. (2021) findings, which revealed a strong association between engagement and the number of hashtags on the Facebook post. The observation implies that tweets with one or two hashtags are associated with a decrease in total engagements for Swahili tweets, irrespective of the engagement type. Perhaps this is because the number of hashtags decreases with multimedia type, as seen in Fig. 2, which as a result, causes less engagement of Swahili tweets. On the other hand, higher numbers of hashtags (four and above) are associated with an increase in engagements, irrespective of the engagement type and language used. However, the magnitudes of impacts are relatively higher for English tweets than that for Swahili tweets. Hashtags seem important in tweets, especially for influencers in social media intending to attract engagement, as reported by Harrigan et al. (2021).

The number of mentions is negatively associated with the engagements, irrespective of either engagement type or language used. This might be because tweets with mentions do not communicate messages to the public but focuses on individual Twitter handles. However, the magnitude of the effect varies per language used and engagement type. According to the results, the number of mentions has a more detrimental effect on Swahili tweets than English tweets. For instance, in the combined analysis (Table 2), while one hashtag is expected to reduce engagements for Swahili tweets by about 84%, the expected reduction is about 63% for English tweets. This might be because Swahili tweets containing mentions are unlikely to have multimedia, reducing engagement

Table 4
Negative Binomial regression results for retweets.

	Overall			English tweets			Swahili tweets		
Variable	Coef	Exp (coef)	P-value	Coef	Exp (coef)	P-value	Coef	Exp (coef)	P-value
Language									
Swahili tweets									
English tweets	0.065	1.07	0.224						
Multimedia type									
No multimedia									
Photo	1.022	2.78	<0.001	0.894	2.45	0.000	0.981	2.67	<0.001
Video	1.842	6.31	<0.001	1.724	5.61	<0.001	1.950	7.03	<0.001
Number of hashtags									
No hashtag									
One hashtag	−0.127	0.88	0.178	0.118	1.13	0.365	−0.331	0.72	0.016
Two hashtags	−0.042	0.96	0.638	0.223	1.25	0.076	−0.433	0.65	0.001
Three hashtags	0.383	1.47	<0.001	0.676	1.97	<0.001	0.088	1.09	0.470
Four hashtags	0.979	2.66	<0.001	1.084	2.96	<0.001	1.012	2.75	<0.001
Five or more hashtags	1.092	2.98	<0.001	1.346	3.84	<0.001	1.072	2.92	<0.001
Number of mentions									
No mention									
One mention	−1.000	0.37	<0.001	−0.520	0.59	<0.001	−1.853	0.16	<0.001
Two mentions	−1.113	0.33	<0.001	−1.174	0.31	<0.001	−1.318	0.27	<0.001
Three or more mentions	−0.709	0.49	0.001	−0.775	0.46	0.013	−0.999	0.37	<0.001
Time of the day									
Early morning (7-9am)									
Late morning (9am-12pm)	0.097	1.10	0.232	0.274	1.31	0.016	0.064	1.07	0.581
Afternoon (12–3pm)	−0.147	0.86	0.074	−0.159	0.85	0.153	−0.030	0.97	0.806
Evening (3–6pm)	−0.446	0.64	<0.001	−0.509	0.60	<0.001	−0.243	0.78	0.019
Nighttime (6pm–7am)	−0.096	0.91	0.419	−0.021	0.98	0.905	−0.020	0.98	0.900
Day of the week									
Weekday									
Weekend	−0.256	0.77	0.000	−0.457	0.63	<0.001	−0.004	1.00	0.971
Tweet year									
2018									
2019	0.020	1.02	0.772	0.088	1.09	0.353	−0.179	0.84	0.075
2020	−0.251	0.78	<0.001	−0.028	0.97	0.768	−0.715	0.49	<0.001
Number of words	0.154	1.17	<0.001	0.143	1.15	0.024	0.127	1.14	0.022
Intercept	−0.011	0.99	0.936	−0.137	0.87	0.551	0.434	1.54	0.018

even more, as observed earlier. Further, the impact of the number of mentions on favorites only (Table 3) and retweets only (Table 4) shows a similar pattern whereby Swahili tweets have fewer engagements than English tweets.

Time of the day has revealed mixed outcomes for English and Swahili tweets, irrespective of the engagement type. The overall analysis that combines both English and Swahili tweets shows that tweets at any other time are associated with low engagements compared to the early morning. Focusing on specific language and engagement type, English tweets are more likely to have a higher number of favorites in the late morning and nighttime, while Swahili tweets are likely to have a lower number of favorites at any time of the day (Table 3). The observation might be due to the time zone difference between English and Swahili-speaking countries. For instance, if a follower is in the United States when it is morning in Tanzania, it is still nighttime in the US. The pattern changes for retweets, whereby only late morning hours are associated with increased retweets for both English and Swahili tweets (Table 4). The findings contradict the ones Harrigan et al. (2021) reported, which predicted citizen engagement with government tweets.

Day of the week is a statistically significant variable for English tweets, irrespective of the engagement type. The results show that English tweets on the weekends are likely to generate less total engagement compared to tweets on the weekdays. This might be because weekdays offer more time to engage tweets, or most people travel during weekdays using Tanzanian airlines; hence more engagement of airlines is evident. The results also show that Swahili tweets are likely to create more weekend engagements than weekdays. However, the results are not statistically significant at a 95% confidence interval.

The year of the tweet revealed mixed outcomes across engagement types and language used. Overall, English tweets are likely to increase total engagements and favorites, while Swahili tweets are less likely to increase engagements. Considering retweets, although the 2019 tweets are associated with increased engagements, they are statistically insignificant at a 95% confidence interval. Contrarily, the 2020 tweets are statistically significantly less likely to generate retweets than the 2018 tweets.

The number of words in a tweet is a statistically significant variable and is associated with increased engagement for Swahili tweets, unlike English tweets. It also shows that the more words in a Swahili tweet, the higher the expected engagement (retweets, favorites, and/or both). This might be because more words give more information, triggering more responses/engagements. The number of words, however, has shown an increased retweet, irrespective of the language used.

The NB regression results provided the overall influence of each variable on the followers' engagements. It was observed that the number of hashtags and tweets with multimedia was associated with high engagements. However, such an approach would not identify the content that drives engagement. Thus, a text network was further applied to understand the key contents associated with followers' engagements. The next section presents the text network results and discussion.

4.2. Text networks analysis results and discussion

In this section, the text networks' results and discussion are presented. The section covers the text network for overall Swahili and English tweets that are tweets with and without hashtags as well as with and without multimedia. The intention was to understand the content of the tweets and the possible reason that might have influenced the engagements. Since the tweets resulted in a large network that cannot be effectively presented, only the 40 most frequent keywords were used for discussion purposes.

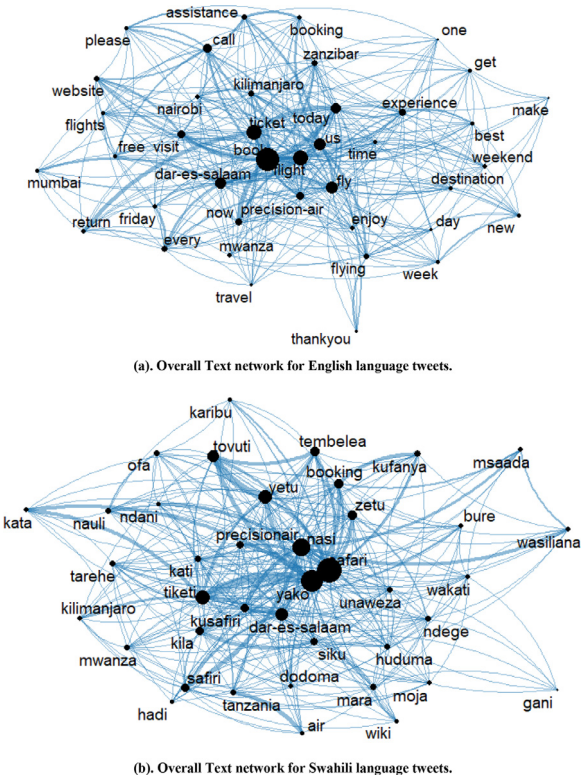

Fig. 3. Overall Text network for English and Swahili language tweets.

4.2.1. Overall text networks of the tweets

Fig. 3(a) and (b) show the text networks for English and Swahili tweets. The two networks revealed several similarities and differences. Both networks show that most of the tweets were for booking advertisements and promotion of the airlines. This is revealed by the keywords *book, ticket, flight,* and *us* for the English tweets network, and *tiketi, yako, safari,* and *nasi* for Swahili tweets network.

The keywords *tiketi, yako, safari,* and *nasi* for Swahili tweets network mean *ticket, yours, flight,* and *us* respectively. Thus, irrespective of the language used, the intention of frequent tweets was to encourage booking the tickets for the upcoming flights. The co-occurred keywords such as *kata tiketi* in Swahili tweets, which means *book tickets*, and *safari nasi,* which means *fly with us* are evident in the English tweets network. Another similarity between the two networks is the customer service support, where clusters like *please, call,* and *assistance* are visible in the English tweets' network, while corresponding keywords *wasiliana, nasi, huduma,* and *tovuti yetu* are observed in the Swahili tweets' network.

Conversely, the two networks have shown some differences. The English tweets network has several destinations that are promoted in the tweets compared to Swahili tweets. In the English tweets network, the destinations such as *Dar-es-salaam, Mwanza, Kilimanjaro, Zanzibar, Nairobi,* and *Mumbai* are observed. On the other hand, only *Dar-es-salaam, Mwanza,* and *Dodoma* are observed in the Swahili tweets network. The presence of *Zanzibar, Nairobi,* and *Mumbai* in the English tweets implies that the tweets' targeted audiences are foreigners and tourists, while *Dar-es-salaam, Mwanza,* and *Dodoma* in Swahili tweets are likely to target local residents.

4.2.2. Text network for tweets with and without hashtags

Fig. 4 shows tweets with and without hashtags for Swahili and the English language. The text networks reveal a couple of similarities and differences. Tweets with hashtags in Fig. 4(a) and (c) show a community where clusters *#tukutaneangani* and *#youarewhywefly* appear to be the key topics. This suggests that most tweets included these hashtags to promote and encourage the air travel experience. Both text networks show these hashtags linked with other clusters such as *book, #25yearsanniversary, experience, us,* and *ticket* for English tweets, and *ticket yako* for Swahili tweets, which encourages booking tickets with the airlines. The

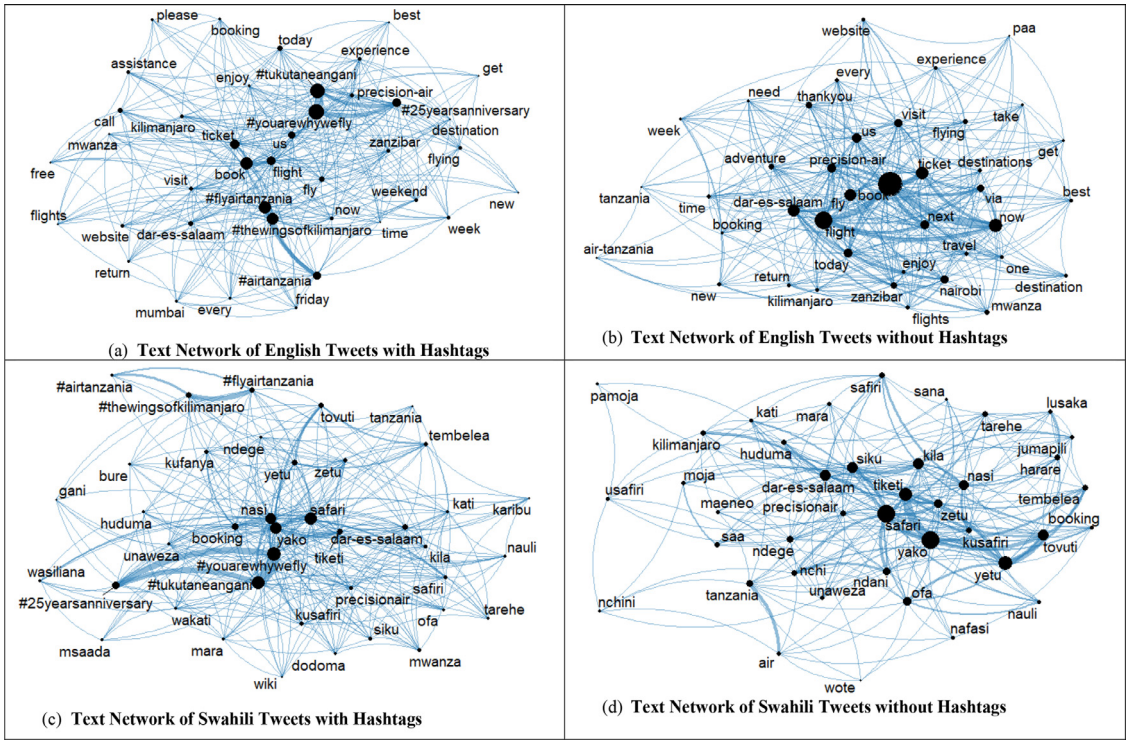

Fig. 4. Text Networks for English and Swahili tweets with and without Hashtags.

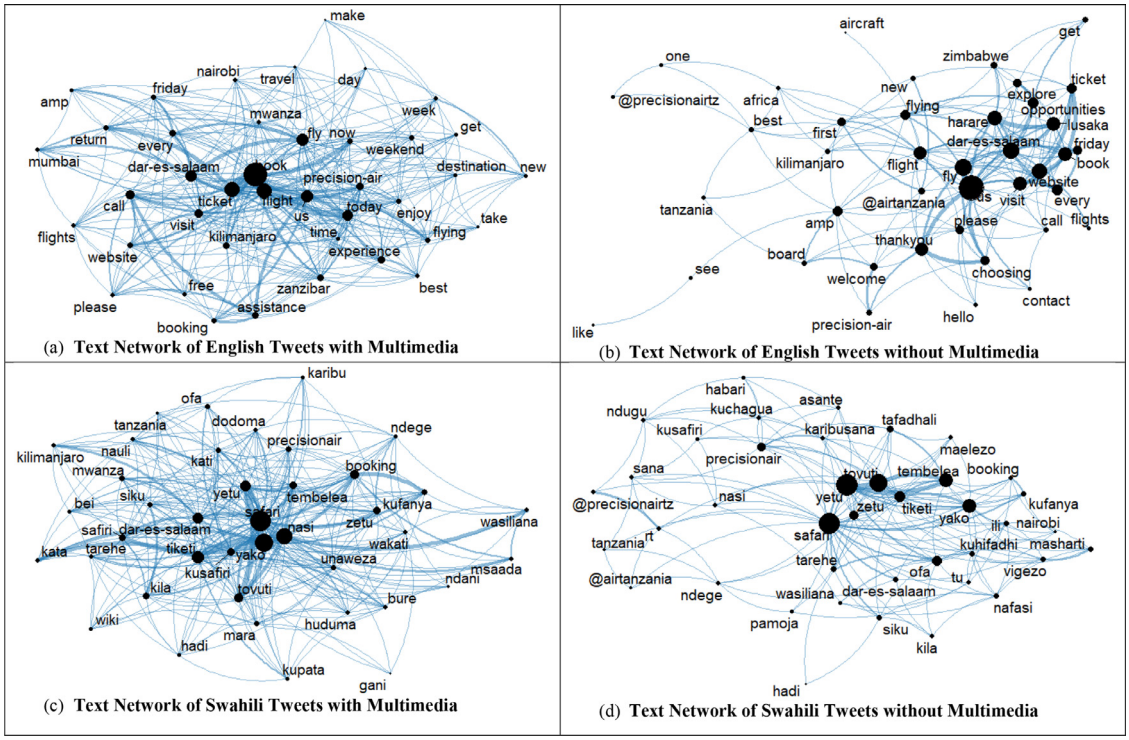

Fig. 5. Text Networks for tweets with and without multimedia.

cluster *precision air* (meaning precision airways) is in the same community as the clusters mentioned above. This suggests that these are marketing hashtags for precision airways and a 25-year anniversary.

Both text networks with hashtags show another community with co-occurred keywords like *#flyairtanzania, #thewings of Kilimanjaro, #airtanzania* suggesting that these are marketing hashtags for Air Tanzania airlines.

The text networks differ slightly in the destinations observed, where *Mwanza* appears in Swahili tweets while *Mumbai* appears in English tweets to target foreigners. Tweets without hashtags on Fig. 4(b) and (d) show a unique observation in which the added destinations, *Lusaka,* and *Harare* as evident in Fig. 4(d). However, other deductions are similar to tweets with hashtags on Fig. 4(a) and (c). This is evident in co-occurred words *safari yako, nasi,* and *ticket* for Swahili tweets and *book, us ticket,* and *flight* for English tweets. This implies the airlines communicate the same message in most of their tweets regardless of the trend. All tweets also show the presence of customer service support.

4.2.3. Text network for tweets with and without multimedia

Fig. 5 shows Text networks with and without multimedia for the English and Swahili tweets. Some similarities and differences have been noted. All networks show that most of the tweets revealed keywords *book, ticket, flight,* and *us* for the English tweets network, and *tiketi, yako, safari,* and *nasi* for the Swahili tweets network. The keywords *tiketi, yako, safari,* and *nasi* for Swahili tweets network mean *ticket, yours, flight,* and *us,* respectively. Thus, irrespective of the multimedia content used, the intention of frequent tweets was to encourage booking tickets for upcoming flights. Another similarity between the networks is the customer service support, where clusters like *please, call,* and *assistance* are visible in the English tweets' network, while corresponding keywords *wasiliana, nasi, huduma,*and *tovuti yetu* are observed in the Swahili tweets' network.

Some differences have also been noted. Tweets without multimedia show development in conversations between customers and customer service. For instance, clusters such as *thank you, choosing, please,* and *hello* in one community for the English language indicate the typical customer care language. Similarly, clusters *kuchagua, karibu sana, maelekezo* and *kila siku* which means *choosing, very welcome, instructions* and *everyday,* respectively, are customer care-related terms in Swahili language tweets. Such words may result from a customer reaction; however, such reactions still seem minimal.

Other notable observations for tweets without multimedia include the clusters *Lusaka, Harare,* and *Zimbabwe* in English language tweets. This is probably a result of widening the market to those specific destinations as they appear to be among the key topics in the network. It is also likely that it was done by Air Tanzania airlines, a cluster in the same community with those destinations. This is seen in Fig. 4, contrary to every other text network in Fig. 5(a), (c), and (d).

4.3. Contributions to literature and implications for practice

This study adds to the body of literature various valuable insights that can be utilized by researchers, business entities, and the public. First, the study evaluated the influence of the language used for communication and promotions on customers/followers engagement. Almost every business entity would be interested to understand if the language used can create more engagement for business growth. The findings from this study provided insights that can be applied to business entities to improve their interactions with customers. For instance, the time of the day has shown contrasting results based on the type of language used. The finding suggests that utilizing different languages at different times of the day would improve customers' engagement. This might be an interesting area to expand on in future studies. Furthermore, the various other tweet contents, such as multimedia, hashtags, and mentions evaluated in this study, can be useful to various business entities to improve their interactions with their customers. Lastly, the methodology (regression-text analysis) applied in this study can be utilized by researchers in other fields to understand not only the association but also the content of consumer reviews using any language other than English and Swahili used in this study.

5. Conclusions and future studies

Understanding tweet characteristics is essential in exploring followers' engagement. A few studies have examined the influence of languages on social media engagement. Further, the approaches used in those studies do not uncover the influence of various tweet contents on consumer engagement. This study uses the hybrid (regression-text) model to uncover what was being discussed in the tweets and the associations between engagements and predictors.

Negative binomial regression analysis is first used to explore both favorites and retweets combined and individually. These are the noted forms of engagement in any tweet. The influences are initially categorized per language used. Characteristics of the tweet are evaluated according to their effect on favorites and retweets. The results for instance, show the number of hashtags and multimedia content of a tweet is associated with high engagements irrespective of the language used. Thus, whichever language the airlines choose in their tweets, multimedia has a tremendous influence on conjuring engagement to their tweets. This aligns with previous findings that multimedia tends to boost tweet popularity.

Furthermore, the study applies a state-of-the-art text mining approach to understand the content of the tweets. It enables us to understand the intentions of the tweet with respect to the language used. It is also cognizant of the tweet characteristics and how they influence engagement. For instance, overall, tweets had the intention to encourage booking tickets for upcoming flights through websites. We also see the presence of customer support. The English language, however, is used to target foreigners and tourists to and from different out-of-country destinations, while the Swahili language is used for, but not limited to, the domestic market. This shows that language is essential in targeting specific groups' engagements.

5.1. Study limitations

The end of this study opens doors to other studies for the areas that could not be covered. It is suggested that more research can be done to understand if tweet engagement has any relation to the performance of airlines, especially African airlines. The key research questions to be answered for this case can be, for instance, whether using certain hashtags or keywords improved customer flow to that airline.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge two anonymous internal reviewers who helped to shape this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Ortiz-Ospina, E. (2019). *The rise of social media - Our World in Data* https://ourworldindata.org/rise-of-social-media (accessed 21, 2022).

Zahoor, S. Z., & Qureshi, I. H. (2017). Social media marketing and brand equity: A literature review. *IUP Journal of Marketing Management, 16*(1).

Ashley, C., & Tuten, T. (2015). Creative strategies in social media marketing: An exploratory study of branded social content and consumer engagement. *Psychology & Marketing, 32*(1), 15–27. 10.1002/mar.20761.

Oliverio, J. (2018). A survey of social media, big data, data mining, and analytics. *Journal of Industrial Integration and Management, 03*(03), Article 1850003. 10.1142/S2424862218500033.

Hanaysha, J. R. (2022). Impact of social media marketing features on consumer's purchase decision in the fast-food industry: Brand trust as a mediator. *International Journal of Information Management Data Insights, 2*(2), Article 100102. 10.1016/J.JJIMEI.2022.100102.

Zeng, B., & Gerritsen, R. (2014). What do we know about social media in tourism? A review. *Tourism Management Perspectives, 10*, 27–36.

Felix, R., Rauschnabel, P. A., & Hinsch, C. (2017). Elements of strategic social media marketing: A holistic framework. *Journal of Business Research, 70*, 118–126. 10.1016/j.jbusres.2016.05.001.

Alalwan, A. A., Rana, N. P., Dwivedi, Y. K., & Algharabat, R. (2017). Social media in marketing: A review and analysis of the existing literature. *Telematics and Informatics, 34*(7), 1177–1190. 10.1016/j.tele.2017.05.008.

Whitelock, J., Cadogan, J. W., Okazaki, S., & Taylor, C. R. (2013). Social media and international advertising: Theoretical challenges and future directions. *International Marketing Review*.

Singh, N., Lehnert, K., & Bostick, K. (2012). Global social media usage: Insights into reaching consumers worldwide. *Thunderbird International Business Review, 54*(5), 683–700. 10.1002/tie.21493.

Eleta, I., & Golbeck, J. (2014). Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior, 41*, 424–432. 10.1016/j.chb.2014.05.005.

Sass, C. A. B., et al., (2020). Exploring social media data to understand consumers' perception of eggs: A multilingual study using Twitter. *Journal of Sensory Studies, 35*(6), e12607. 10.1111/joss.12607.

Lim, H., & Fussell, S. (2017). Understanding how people attend to and engage with foreign language posts in multilingual newsfeeds. *Proceedings of the International AAAI Conference on Web and Social Media, 11*(1) SE-Poster Papers.

Jung, S. H., & Jeong, Y. J. (2020). Twitter data analytical methodology development for prediction of start-up firms' social media marketing level. *Technology and Society, 63*, Article 101409. 10.1016/j.techsoc.2020.101409.

Mazumdar, S., & Thakker, D. (2020). Citizen science on Twitter: Using data analytics to understand conversations and networks. *Future Internet, 12*(12). 10.3390/fi12120210.

Aswani, R., Kar, A. K., Aggarwal, S., & Vigneswara Ilavarsan, P. (2017). Exploring content virality in facebook: A semantic based approach. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): 10595* (pp. 209–220). LNCS. 10.1007/978-3-319-68557-1_19.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika, 70*(1), 269–274. Accessed: 02, 2017. [Online]. Available https://www.nuffield.ox.ac.uk/users/cox/cox155.pdf .

Kutela, B., & Teng, H. (2019). The influence of campus characteristics, temporal factors, and weather events on campuses-related daily bike-share trips. *Journal of Transport Geography, 78*, 160–169. 10.1016/j.jtrangeo.2019.06.002.

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed poisson, and negative binomial models. *Psychological Bulletin, 118*(3), 392–404. 10.1037/0033-2909.118.3.392.

Booth, J. G., Casella, G., Friedl, H., & Hobert, J. P. (2003). Negative binomial loglinear mixed models. *Statistical Modeling, 3*, 179–191. Accessed30, 2017[Online]. Available http://journals.sagepub.com/doi/pdf/10.1191/1471082X03st058oa .

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *Journal of High Technology Management Research, 15*(1), 37–50. 10.1016/j.hitech.2003.09.003.

Kushwaha, A. K., Kar, A. K., & Dwivedi, Y. K. (2021). Applications of big data in emerging management disciplines: A literature review using text mining. *International Journal of Information Management Data Insights, 1*(2), Article 100017. 10.1016/J.JJIMEI.2021.100017.

Benoit, K., et al., (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software, 3*(30). 10.21105/joss.00774.

Kutela, B., Kitali, A. E., & Kidando, E. (2022). Examining the influence of alternative fuels' regulations and incentives on electric-vehicle acquisition. In *International conference on transportation and development* (pp. 196–206). 10.1061/9780784484340.018.

Saheb, T., Amini, B., & Kiaei Alamdari, F. (2021). Quantitative analysis of the development of digital marketing field: Bibliometric analysis and network mapping. *International Journal of Information Management Data Insights, 1*(2), Article 100018. 10.1016/J.JJIMEI.2021.100018.

Hunter, S. (2014). A novel method of network text analysis. *Open Journal of Modern Linguistics, 4*, 350–366. 10.4236/ojml.2014.42028.

Jiang, C., Bhat, C. R., & Lam, W. H. K. (2020). A bibliometric overview of Transportation Research Part B: Methodological in the past forty years (1979–2019). *Transportation Research Part B: Methodological, 138*, 268–291. 10.1016/j.trb.2020.05.016.

Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights, 1*(1), Article 100008. 10.1016/J.JJIMEI.2021.100008.

Y. Kim and S.-N. Jang, "Mapping the knowledge structure of frailty in journal articles by text network analysis," 2018, doi: 10.1371/journal.pone.0196104.

Kutela, B., Novat, N., Adanu, E. K., Kidando, E., & Langa, N. (2022). Analysis of residents' stated preferences of shared micro-mobility devices using regression-text mining approach. *Transportation Planning and Technology, 45*(2), 159–178. 10.1080/03081060.2022.2089145.

Kutela, B., Novat, N., & Langa, N. (2021). Exploring geographical distribution of transportation research themes related to COVID-19 using text network approach. *Sustainable Cities and Society, 67*. 10.1016/j.scs.2021.102729.

Paranyushkin, D. (2011). Identifying the pathways for meaning circulation using text network analysis. *Venture Fiction Practice, 2*(4). Accessed: 27, 2020. [Online]Available www.noduslabs.com .

Kutela, B., Langa, N., Mwende, S., Kidando, E., Kitali, A. E., & Bansal, P. (2021). A text mining approach to elicit public perception of bike-sharing systems. *Travel Behaviour and Society, 24*, 113–123. 10.1016/j.tbs.2021.03.002.

Kutela, B., Das, S., & Dadashova, B. (2021). Mining patterns of autonomous vehicle crashes involving vulnerable road users to understand the associated factors. *Accident Analysis and Prevention*, Article 106473. 10.1016/J.AAP.2021.106473.

Csárdi, G. (2020). *'igraph'; Network analysis and visualization* Accessed: 01, 2020. [Online]Available https://igraph.org/r/ .

Anand, K., Urolagin, S., & Mishra, R. K. (2021). How does hand gestures in videos impact social media engagement - Insights based on deep learning. *International Journal of Information Management Data Insights, 1*(2), Article 100036. 10.1016/J.JJIMEI.2021.100036.

Ahn, J., Son, H., & Chung, A. D. (2021). Understanding public engagement on twitter using topic modeling: The 2019 Ridgecrest earthquake case. *International Journal of Information Management Data Insights, 1*(2), Article 100033. 10.1016/J.JJIMEI.2021.100033.

Harrigan, P., Daly, T. M., Coussement, K., Lee, J. A., Soutar, G. N., & Evers, U. (2021). Identifying influencers on social media. *International Journal of Information Management, 56*, Article 102246. 10.1016/J.IJINFOMGT.2020.102246.