DOI: http://doi.org/ 10.4038/tare.v21i1-2.5459

Tropical Agricultural Research & Extension 21 (1 & 2): 2018

TWITTER™ ON AQUACULTURE: UNDERSTANDING THE LATENT INFORMATION USING R

Tharindu Bandara[1]*, K Radampola[2]

[1]Faculty of Biosciences and Aquaculture, Nord University, Norway

[2]Department of Fisheries and Aquaculture, Faculty of Fisheries and Marine Science and Technology, University of Ruhuna, Matara

ABSTRACT

Social media networks (Twitter™, Facebook™) have significant importance in sharing knowledge and ideas among people. Data mining in these platforms provides valuable information for scholarly use in various fields of agriculture and aquaculture. The purpose of this study was to understand the latent information of twitter messages (tweets) related to the aquaculture. R programming language and the *TwitteR* package were used to extract and analyze the tweets (n=500). The Topic modeling approach was used to identify the key aquaculture themes that can be used to classify the tweets. Descriptive analysis of tweets indicated that Twitter users have used 17 language profiles. 372 twitter profiles have tweeted about aquaculture. Europe and North America collectively had the highest number of tweets (60%). "GAA_Aquaculture" (2.2%), "Farming Tilapia" (1.8%), "GrowAquaponics" (1.6%), "Wild4salmon" (1.2%) and "FAOfish" (1.2%) were top twitter profiles with the highest number of tweets. Term 'salmon' was significantly correlated ($p<0.05$) with 'Wild salmon', 'bute fish', 'Argyll' and 'fish farm get out'. Results of the Topic model classified the tweets into five key themes (Food security and sustainable aquaculture, fish nutrition, sea lice infestation in salmon aquaculture and Tilapia aquaculture). These results indicated that mining Twitter data can be effectively used for understanding the latent information about aquaculture.

Key Words: Twitter, Aquaculture, R programming, Data mining, Topic modeling, Social-media

INTRODUCTION

Twitter™ - a popular micro-blogging social media platform intended to express thoughts, status, and ideas among its users. Twitter allows sending brief text messages (each consists of 140 characters named as tweets), and small micro media. By the second quarter of 2017, it had 328 million active users indicating that one of the largest social media networks in the world (Statista, 2017b). Apart from general usage as a social media platform, twitter data provide invaluable information for scholarly use. These include dissemination of the research findings, identifying research topics (as a secondary data source) and conducting surveys (Adolphus, 2017).

During the last decades, advances in data mining techniques have significantly contributed to the discovering of new knowledge (Aggarwal and Zhai, 2012). In this regard, information retrieval by text mining approaches has gained increasing popularity among the researchers. Text mining can be defined as the automated processing of a large amount of textual data for the purposes of meaningful analysis and interpretation (Reilly Jr, 2009) Applications of text mining approaches can be seen in the fields of agriculture (Prabhakar *et al*., 2010) Journalism (Junqué de Fortuny *et al*., 2012) and literature (Nuzzo *et al*., 2010). With the aid of open source programming languages (R, Python, Scala) extracting text data from the tweets has become more flexible. In this context, readily available large number of packages and strong community support (e.g. stackoverflow.com) enable R as an ideal source for analyzing Twitter data.

*Corresponding author: tharinduacademia@hotmail.com

Topic modeling approach is widely used among the academic community for extracting the latent information of tweets (Alvarez-Melis and Saveski, 2016). In scholarly usage, discovering the abstracted topics hidden in each tweet is usually achieved by topic modeling function: Latent Dirichlet Allocation (LDA). LDA classifies the text documents/tweets into major topics by an automated process. A formal mathematical approach to topic modeling has been presented by Wallach (2006). However, informal and more user-friendly explanations also presented by Welleck (2014) and Weingart (2011).

As the fastest growing food production sector in the world, aquaculture provides a significant contribution to the global food security (FAO, 2003). Beyond 2030, aquaculture would be a dominant fish supplier in the world (World Bank, 2013). Despite these, as the other food production sectors in the world, global aquaculture is also finding solutions for various issues related to it such as sustainability of feed resources, issues in fish health management and environmental pollution. Apart from traditional journals, communications of these research findings through twitter network is common in the era of Web 2.0 (Ortega, 2017). However, these tweets are largely unstructured and finding the latent topics in tweets of particular interest need careful text mining approach. Therefore, the objective of the present study was to understand the latent information/topics of aquaculture related tweets employing machine learning approach.

MATERIALS AND METHODS

Extracting tweets from the Twitter database was initiated by creating the Twitter application at Twitter developer's platform (https://dev.twitter.com). Twitter developer's platform allows users to obtain relevant security credentials (e.g. consumer key, consumer API) to interact with twitter API. *TwitteR* package from CRAN repository was obtained and installed in R statistical program to search, extract and carry out advanced functions on tweets. Twitter authentication with R was done by using the consumer key, consumer API, access token and access token secret provided by the created application. After the connection has established, *searchTwitter()* function in *TwitteR* with keyword *#aquaculture* was used to find out tweets (n=500) about aquaculture. Extracted tweet list was transformed into a data frame by using *twListoDF()* function. Subsequently, metadata analysis of tweets was performed to understand the language profiles and geographical distribution of the tweets. For further statistical analysis and to follow up linguistic rules, data frame was converted into a text corpus. In the text corpus, extracted tweets were cleared for URLs, stop words and various punctuation marks followed by lower case transformation of letters. Cleaning of the tweets for above contents was done by using R packages *tm* and *gsub*. Transformation of the text corpus into document term matrix was done for further analysis of texts. Most frequent words from all tweets were then identified and plotted by using *ggplot2* and *wordcloud* packages. *findAssocs* function from *tm* package was used to identify closely related words.

To classify the tweets into the major topics, the *topicmodels* package from CRAN repository was used. LDA function from the topic model package was used to evaluate the number of topics and key terms associated with each topic. Based on the key terms of each topic, topics were assigned to relevant aquaculture theme. Subsequently, by using the LDA function, each of tweet was classified into relevant aquaculture theme with the highest probability.

RESULTS AND DISCUSSION

Metadata analysis of tweets indicated that Twitter users have used different languages in their profiles. 78.8 % of the Twitter users have used English as their primary language. Other top languages included French (10.4%) and Spanish (5.8%) (Table 1). However, content analysis of tweets indicates that every tweet has written in English. Location settings of the observed twitter profiles indicated that 24.8%

of tweets originated from Europe followed by North America (39.2%) (Figure 1). Europe and North America are the largest aquaculture production areas in the world (FAO 2016). Unlike in other parts of the world (e.g. Asia, Africa), aquaculture production of these continents depends on the high level of technology and better e-infrastructure. Major aquaculture firm and institutes in these continents use Twitter as one of the powerful mass communication media (e.g. @MHCanada: Marine Harvest Canada, one of the largest salmon production companies in Canada). Moreover, Twitter had a significant share (53.3%) of active users within these two continents by the end of 2017 (Statista, 2017a). Improved e-communication and a larger number of active users might be a reason for the higher number of aquaculture related tweets circulating within these continents. More importantly, location unidentified tweet percentage was 18.8%. Enabling location services is turned off by default on Twitter (Twitter, 2017). This situation leaves most of the tweets on the Twitter network as non-georeferenced tweets (Leetaru *et al*., 2013).

Table 1: Default Language profiles of the tweets as percentage (n=500)

Language	Percentage of tweets (%)
English	78.8
French	10.4
Spanish	5.8
German	0.8
Turkish	0.8
Norwegian	0.6
Finnish	0.4
Italian	0.4
Portuguese	0.4
Swedish	0.4
Arabic	0.2
Indonesian	0.2
Dutch	0.2
Polish	0.2
Russian	0.2
Ukrainian	0.2

Analysis of Twitter profiles indicated that 372 twitter profiles tweeted about aquaculture. 'GAA_Aquaculture', 'FarmingTilapia', 'GrowAquaponics', 'Wild4Salmon' and 'FAOfish' were top twitter accounts based on the frequency of tweets (Figure 2). GAA_Aquaculture (Global aquaculture alliance) is one of the largest aquaculture alliances in the world, responsible for sustainable aquaculture and disseminating the knowledge on aquaculture. FAOfish is the official twitter account for fisheries and aquaculture department of United Nation's food and agricultural organization. FAOfish, disseminates various news, information, conference information about aquaculture under different aquaculture themes. Twitter account 'Wild4Salmon' is hosted by Michelle Young, Canadian environmentalist concerns about aquaculture, wild salmon, and marine environment. Wild4Salmon was followed by a significant number of followers (2636 as per 20/9/2017) interested in her activities. 'Farming Tilapia 'disseminates information about Tilapia aquaculture. 'Grow Aquaponics' is USA based

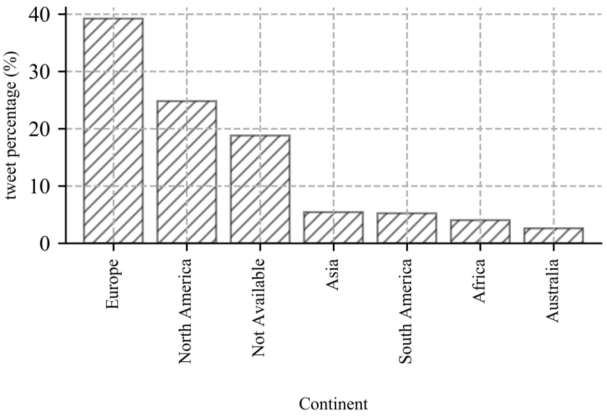

Figure 1: Percentage of tweets by different locations

commercial aquaculture firm to disseminate information about aquaculture technology as aquaponics systems.

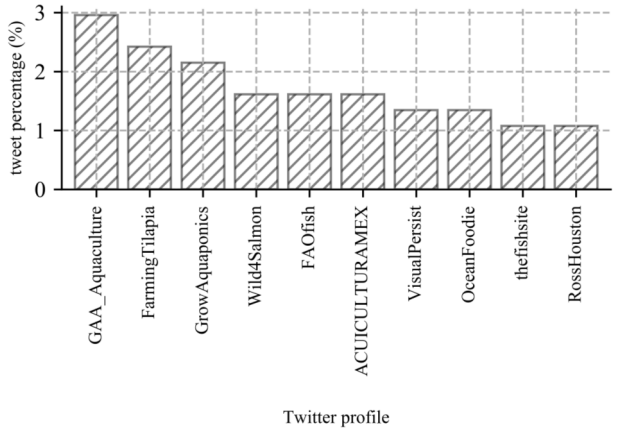

Figure 2: Top 10 twitter accounts tweeted on aquaculture

Term frequency analysis of the text corpus indicated that 'aquaculture', 'salmon', 'fish', 'sustainable' and 'lice' were most frequent keywords (Figure 3). Word association results indicated that keyword 'salmon' was correlated with 18 keywords (Figure 4). Significant correlation ($p<0.05$) was found between the terms salmon- wild salmon, salmon- fish farms getout, and salmon- Argyll and salmon- bute fish. Isle of bute and Argyll is one of the important sites for salmon industry in Scottish Highlands. In 2017, Argyll and Bute salmon industry have earned significant earning (£ 120 million) compared with other aquaculture practices in the region (Ferguson *et al*., 2016). "Fish farms get out "is popular hashtag (#) among the individuals in social movements against the unsustainable salmon farming around the world. On the other hand, keywords including 'wild salmon', 'lice', 'plagues' 'parasite' are representing the current issues in salmonid aquaculture (Morton and Routledge 2016). In today's context, escaped wild salmon has considerable effect on the wild salmon populations. Sea lice infestation from the intensive salmonid farming is one of the greatest threats to both commercial salmonid farming and wild salmonids (Torrissen *et al*., 2013). These results indicated that there is considerable discussion in twitter community on the salmon farming industry.

Topic modeling function has classified the text corpus into five major topics. Each of the topic consisted of six keywords (Table 2).

Based on the top keywords represent by each topic, Topic modeling approach has explained the five key areas of the aquaculture. Probability distribution of the topics in each tweet showed that all five themes existed in each tweet by varied proportion. However, based

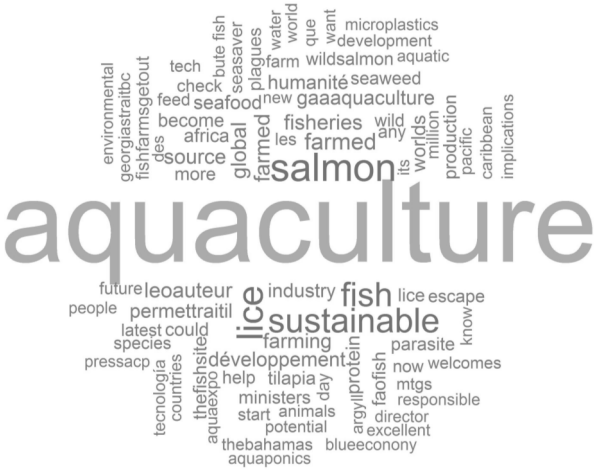

Figure 3: Word cloud representation of the most frequent keywords within the text corpus (n=100)

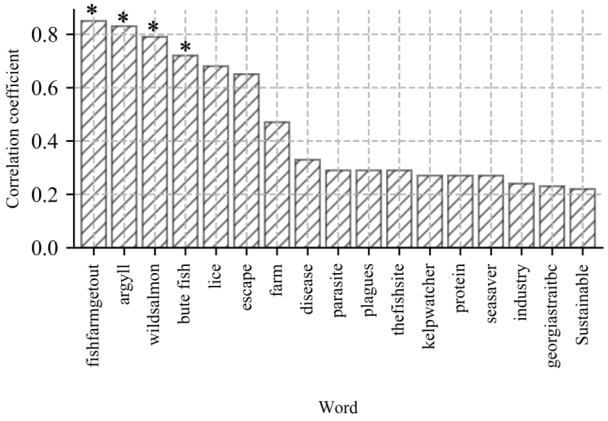

Figure 4: Words that are correlated with the word-salmon (note that words marked with * significantly correlated ($p<0.05$) with word salmon)

upon the highest probable topic in each tweet, relevant tweet has classified to one of the key themes by topic model function (e.g. tweet 1: fish nutrition, tweet 2: Aquaculture technology). For entire corpus, 19.8% tweets were classified into the food security and aquaculture theme, 26% tweets into fish nutrition, 20.4% tweets into sea lice infestation and salmon aquaculture, 17.8% tweets into aquaculture technology and 16% into Tilapia aquaculture by topic model. Sustainability issues of fish meal have driven world aquaculture practices for alternative feed ingredients. Every year, a number of alternative ingredients are tested as a replacement for fish meal. Findings about these ingredients are communicated by various institutes and aquaculture firms through social media. This may create a significant discussion in twitter community about aquaculture nutrition. Key themes including aquaculture technology, Tilapia aquaculture and food security have been identified as the important drivers for achieving sustainable development goals in 2030 agenda by FAO (FAO, 2017). Therefore, further updates on new knowledge related to the above topics may create significant community participation in aquaculture. Moreover, finding possible solutions to issues raised by various aquaculture topics might be important for future growth of aquaculture practices.

CONCLUSION

Social networks are disseminating a large amount of information every day. Application of data mining techniques to understand the latent information on social networks is important for future research and decision-making. Findings of the present research indicated that R can be used to mine the data on aquaculture in twitter network. It has shown that twitter community is largely discussing various topics in aquaculture and possible issues related to it. Further research on other text mining approaches with using different tools in larger data set may provide more latent topics on aquaculture.

REFERENCES

Adolphus M 2017 How to use Twitter for academic research. Available at: http:// www. emeraldgrouppublishing.com/ research/guides/ management/ twitter.htm (Accessed: 12 November 2017)

Aggarwal CC, Zhai CX 2012 A survey of text clustering algorithms. In:Mining Text Data. Springer, New York. pp. 77 -128

Alvarez-Melis D, Saveski M 2016 Topic Modeling in Twitter: Aggregating Tweets by Conversations. Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016), California, USA. 519-522.

FAO 2003 The Role of Aquaculture in Improving Food Security and Nutrition. Food and Agricultural Organization of

Table 2: Classification of the tweets based on the LDA function

Topic	Key terms as defined by the model	Assigned topic based on the key terms
Topic 1	Aquaculture, salmon, protein, sustainable, security, farm	Food security and sustainable aquaculture
Topic 2	Aquaculture, Sustainable, fish, feed, fish oil	Fish Nutrition
Topic 3	Lumpfish, salmon, lice, plagues, parasite, industry	Sea lice infestation in salmon aquaculture
Topic 4	Cages, aquaponics, Technology, RAS, fish, vaccines	Aquaculture technology
Topic 5	Development, Tilapia, Africa, Fisheries, fish, Sub-Sahara	Tilapia Aquaculture

the United Nations. Rome. Avialble at: http://www.fao.org/docrep/MEETING/006/Y8871e.HTM (Accessed: 10 November 2017)

FAO 2016 Global Production Statistics. Avialable at: http://www.fao.org/fishery/statistics/global-production/query/en (Accessed: 16 October 2017)

FAO 2017 The 2030 Agenda and the Sustainable Development Goals: The Challange for Aquaculture Development and Management. Food and Agricultural Organization of the United Nations. Rome.

Ferguson N, Alistair B, Craig C, Forteith J, Francis N, Jurgensen I, Macleod J, Mcconnachie J, Loudan S, Morton CA 2016 Argyll and Bute Economic forum Report. Argyll and Bute Council. Argyll. pp. 35-36.

Junqué de Fortuny E, De Smedt T, Martens D, Daelemans W 2012 Media coverage in times of political crisis: A text mining approach. Expert Syst. Appl. 39:11616-11622.

Leetaru K, Wang S, Cao G, Padmanabhan A, Shook E 2013 Mapping the global Twitter heartbeat: The geography of Twitter. First Monday. doi: 10.5210/fm.v.18i5.4366

Morton A, Routledge R 2016 Risk and precaution: Salmon farming. Mar. Policy 74:205-212.

Nuzzo A, Mulas F, Gabetta M, Arbustini E, Zupan B, Larizza C, Bellazzi R 2010 Text Mining approaches for automated literature knowledge extraction and representation. Stud. Health Technol. Inform. 160:954-8.

Ortega JS 2017 The presensece of academic journals on Twitter and its relationship with dissemination (tweets) and research impact (citations). Aslib Journal of Information Management. 69: 674-687.

Prabhakar TV, Neelam LK, Balaji V 2010 Agrotags: A contribution towards improved digital information management in agricultural research. Ann. Libr. Inf. Stud 57:278-281.

Reilly Jr BF 2009 When Machines Do Research: Automated Analysis of News and Other Primary Source texts. Jour. Lib.Adm. 49:507-517

Statista 2017a Distribution of Twitter users worldwide from 2012 to 2018, by region: Available at: https://www.statista.com/statistics/303684/regional-twitter-user-distribution/ (Accessed: 01 December 2017).

Statista 2017b Number of monthly active Twitter users worldwide 1st quarter 2010 to 2nd quarter 2017 (in millions). Available at: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ (Accessed: 06 September 2017)

Torrissen O, Jones S, Asche F, Guttormsen A, Skilbrei OT, Nilsen F, Horsberg TE, Jackson D 2013 Salmon lice – impact on wild salmonids and salmon aquaculture. J. Fish. Dis. 36:171-194.

Twitter 2017 FAQs about adding location to your tweets Available at: https://support.twitter.com/articles/78525 (Accessed: 24 September 2017)

Wallach HM 2006 Topic modeling: beyond bag-of-words. Proceedings of the 23rd international conference on Machine learning. Pittsburgh, Pennsylvania, USA: ACM. pp. 977-984.

Weingart S 2017 Topic modeling and network analysis. Available at: http://www.scottbot.net/HIAL/index.html@p=221.html (Accessed: 27 June 2017)

Welleck S 2017 These are your tweets on LDA (Part I). Available at: https://wellecks.wordpress.com/2014/09/03/these-are-your-tweets-on-lda-part-i/ (Accessed: 15 June 2017)

World Bank 2013 FISH TO 2030, Prospects for Fisheries and Aquaculture. World Bank. Washington. pp.5-6.