FUDMA Journal of Sciences (FJS)
ISSN: 2616-1370
Vol. 2 No. 2, June, 2018, pp 14-20

SENTIMENT ANALYSIS ON TWITTER HEALTH NEWS

[1]Kolajo, T. and [2]Kolajo, J. O.

Department of Computer Science, Federal University Lokoja, Nigeria
Department of Information and Communication Technology, National Health Insurance Scheme, Abuja
Email: taiwo.kolajo@fulokoja.edu.ng, jkolajo@nhis.gov.ng

ABSTRACT
Microblogging has become a generally accepted way of expressing opinions and sentiments about products, services, media, institutions to mention but few. A lot of research has focused on analyzing Twitter health news for topic modelling using various clustering approaches, but few have reported it for sentiment analysis. The fact that such data contains potential information for revealing the opinion of people about health services and behaviours make it an interesting study. In this paper, general sentiments about Twitter health news was investigated. Natural language processing and text mining tool, AYLIEN API was used to extract sentiments subjectivities and polarities from a previously uncategorized dataset. The result shows that most of the tweets in Twitter health news are objective, that is, expressing facts with an average of 64% objective while 34% are personal views or opinions (subjective) and subjectivity confidence of 0.9. Sentiment polarity reveals 9% positive, 19% negative and 72% neutral with polarity confidence of 0.6.

Keywords*: Sentiment analysis, Twitter health news, AYLIEN API, Sentiment Polarity and Subjectivity*

INTRODUCTION

Use of user-generated content for health studies is gradually increasing (Yoon, Elhadad, and Bakken, 2013). As at first quarter of 2018, Twitter monthly users have grown to 336 million who generate above 500 million of 280-character tweets per day (Statista, 2018). Twitter is the most prolific provider of social media research data (Ottoni, et al., 2014). Twitter is one of the most visited sites making it a suitable platform to analyse information spread and sentiments (Godea *et al*., 2015). The abundance of information available in social media and health news forum along with free and rich expression of opinions has attracted public health community to analyse the reaction of people to health services and behaviours. Real-world health services and behaviours can be promoted through the use of Twitter health news. Unlike most of the ecologic assessment methods which capture and allow individual to report current location, activity and social surrounding at any particular period (Schwartz and Stone, 1998), tweets are not specific intermittent stimulus dependent as they represent more naturalistic content with the availability in large volume as an added advantage (Yoon *et al.*, 2013).

Performing content mining analysis on social media data in order to study people's opinion about health services and behaviours is important because gaining full understanding of such opinions from text has been difficult as a result of their complexity. Web content mining focuses on discovering meaningful knowledge (e.g. topics, sentiments) from data such as blogs, online mailing lists, and social media by applying various techniques such as machine learning, data mining, information retrieval, natural language processing and statistics (Nasraoui, 2008).

The process of automatically measuring the speculation, emotions, evaluations, and opinion expressed in a text is referred to as sentiment analysis (Turney, 2002; Pang and Lee, 2008; Liu, 2010, 2012; Kranjc, *et al*., 2015). The speculation, emotions, evaluations, and opinion can either be negative, positive or neutral. Sentiment analysis has been applied to news article, social media post, product reviews, emails, biographies, narratives, and fairy tales, disease outbreak, online health communities (Ji, Chun, and Geller, 2013; Greaves *et al.*, 2013; Korkontzelos *et al*., 2016). Sentiment analysis is a challenging task especially in the context of user-generated content due to informal usage of terms, noisy data, incomplete statements, slangs, irregular sentences, statements with grammatical and spelling errors, abbreviated words, and comments with improper sentence structure that are prevalent in social media content (Petz *et al*., 2012; Panagiotou, Katakis, and Gunopulos, 2016; Katragadda, Benton, and Raghavan, 2017).

In this paper, natural language processing and AYLIEN API, a novel text mining API were employed to extract sentiment polarities and subjectivities expressed by Twitter users towards health news. The rest of the paper is structured as follows. The next section presents the related works followed by materials and methods used in the research work. Thereafter, we have result and discussion section followed by conclusion and further work.

RELATED WORKS

Sentiment analysis is an active research area and it is being used extensively for analyzing and summarizing opinions of social media users (Thakkar and Patel, 2015). Many research works have been done in relation to sentiment analysis. This section presents sentiment analysis research works.

Prieto *et al.* (2014) used "pregnancy", "flu", "depression", and "eating disorders" as query terms to collect tweets and applied regular expression and machine learning algorithms to monitor public opinion and disease information in Spain and Portugal. They achieved a promising result compared to baseline methods with F-measure values of 0.8 and 0.9. A supervised learning approach to extracting sentiments and information dissemination from tweets was also proposed by Godea *et al.* (2015). However, their systems depend on training labelled set.

A unified approach for sentiment analysis was proposed by Ribeiro, Weigang, and Li (2015). The approach is made up of four modules which are: data collection, noise reduction, lexicon generation, and sentiment classification. The proposed approach was experimented using iPhone 6 dataset which shows significant improvement when compared with similar methods. However, it suffers in the face of a large dataset.

Sentiment analysis on tweets about diabetes using an aspect-level approach was conducted by Salas-Zarate, *et al*. (2017). The sentiment was calculated using N-gram methods (N-gram before, N-gram around, and N-gram after). The experimental results show that N-gram around method performed best with a precision, a recall, and an F-measure of 81.93%, 81.13%, and 81.24% respectively.

An hybrid approach has also been used for sentiment analysis. Al-Amrani, Lazaar, and El-Kadiri (2018) combined random forest and support vector machine to extract sentiment from product review dataset. In the same vein, Hassan *et al.* (2018) also adopted a hybrid approach for sentiment analysis. The authors concluded that hybrid approach performed better than individual algorithms.

In contrast, this work aims at extracting sentiments subjectivities and polarities from a previously uncategorized dataset using natural language processing and text mining techniques.

MATERIALS AND METHODS

This section presents the materials and methodology workflow adopted for this work which includes three modules; data collection, preprocessing and analysis. The methodology workflow is depicted in figure 1.

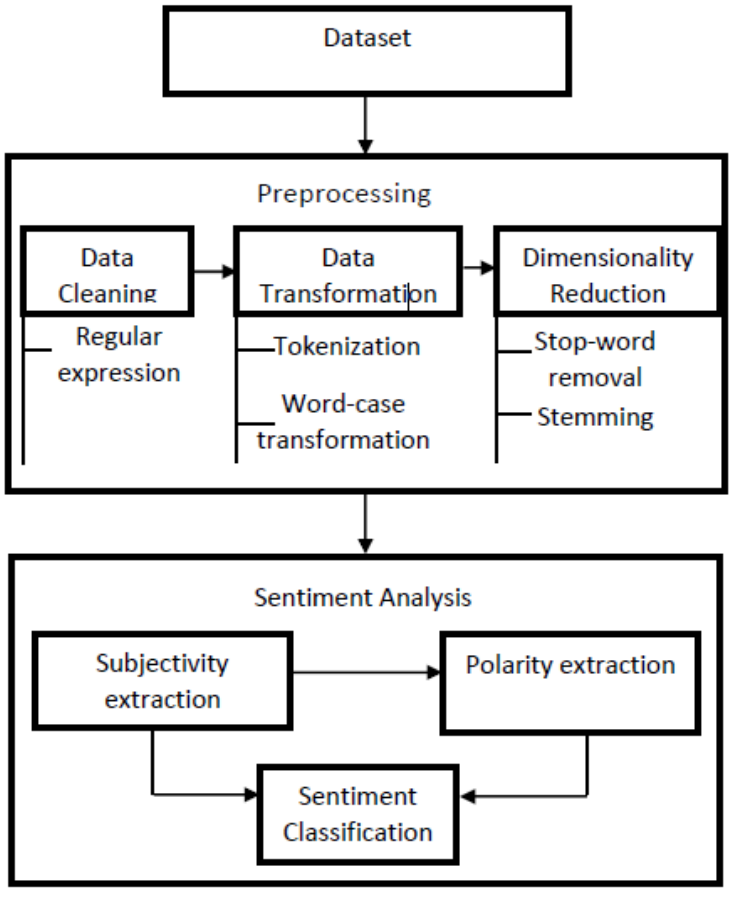

Fig. 1: Sentiment Analysis Methodology Workflow

Twitter Health News Data Collection

The dataset for this paper was collected from health news tweet dataset. The dataset comprises of 16 various health-related news gathered from August 2011 to December 2014. The dataset includes foxnewshealth, everydayhealth, cbchealth, latimeshealth, kaiserhealthnews, goodhealth, nprhealth, NBChealth, msnhealthnews, cnnhealth, bbchealth, usnewshealth, reuters_health, gdnhealthcare, nytimeshealth, and wsjhealth. 7 out of the 16-health related news dataset were selected for analysis. The selected health-related news were bbchealth, cbchealth, cnnhealth, eerydayhealth, gdnhealthcare, msnhealthnews, and usnewshealth. Each of the selection was extracted and converted to CSV format for easy analysis. A total of 24,588 tweets were analysed in all.

Twitter Health News Preprocessing

The preprocessing steps include tweet cleaning, transformation, and dimensionality reduction. Unlike carefully authored news, most of the information on social media platforms contains informal usage of terms, noisy data, incomplete statements, slangs, irregular sentences, statements with grammatical and spelling errors, abbreviated words, and comments with improper sentence structure (Panagiotou *et al.*, 2016; Katragadda *et al.*, 2017). Thus, it is necessary to remove non-ASCII characters, hashtags, URLs, name mentions, punctuation, emoticons, symbols, and retweet symbols prior to analysis. Text cleaning was achieved using a regular expression.

In Tweet transformation stage, tokenization was performed to represent tweets content as a vector of feature. Thereafter, all cases were transformed to lower cases. Both tokenization and word-case transformation were done using "tokenize" and "transform cases" operators in RapidMiner.

Since processing a large dataset with high dimension algorithmically is more difficult, reducing the dimensionality of such dataset becomes necessary. Two methods of dimensionality reduction that were applied were stop-word removal and stemming using a stop-word dictionary and WordNet respectively.

Twitter Health News Sentiment Analysis

The preprocessed tweets are analysed to discover sentiment about twitter health news at this stage. There has been more research in machine learning and computational linguistics about automated detection of attitudes and opinions in text (Yoon *et al.*, 2013). Through the application of computational analytic techniques, sentiment analysis or opinion mining provides valuable insight about the author's perspective or emotion about a document or topic by revealing whether the tone is negative, positive or neutral and whether the text is objective. (i.e. expressing a factual information about the world) or subjective (i.e. reflecting author's feelings or beliefs). In this paper, text analysis tool, AYLIEN API was used to extract sentiments from twitter health news dataset. AYLIEN API is a package of information retrieval, natural language processing and machine learning used for the extraction of meaning and insight from textual or visual content. It can be used as an extension from within RapidMiner, made up of a suite of operators such as sentiment analysis, language detection, entity extraction, related phrases, and hashtag suggestion which gives room for easy analysis.

RESULT AND DISCUSSION

This section presents the results of the study along with the discussion of the results. The result from seven different Twitter health news is presented in Tables 1 and 2. Thereafter, to present the result as a whole, the total results from the seven different Twitter health news are combined and the polarity, subjectivity, and Polarity/Subjectivity Confidence are depicted in figures 2-4.

Table 1: Sentiment Polarity Analysis of Twitter Health News

Health News	Polarity			Polarity Confidence (Average)
	Positive (%)	Negative (%)	Neutral (%)	
bbchealth	3	16	81	0.6
cbchealth	3	20	77	0.7
cnnnealth	12	20	68	0.6
everydayhealth	16	19	65	0.6
gdnhealthcare	11	19	70	0.7
msnhelthnews	3	19	78	0.7
usnewshealth	16	19	65	0.6

While most of the Twitter health news (from August 2011 to December 2014) is neutral in their opinion, the negative sentiment still carries higher percentage than that of positive sentiment. This implies that negative sentiment is expressed more than positive sentiment in Twitter health news.

Table 2: Sentiment Subjectivity Analysis of Twitter Health News

Health News	Subjectivity		Subjectivity Confidence (Average)
	Objective (%)	Subjective (%)	
Bbchealth	69	31	0.9
cbchealth	70	30	0.9
cnnnealth	65	35	0.9
everydayhealth	55	45	0.9
gdnhealthcare	66	34	0.9
msnhelthnews	67	33	0.9
usnewshealth	56	44	0.9

With a higher percentage of objective and high subjectivity confidence reveals that most users in Twitter health news are expressing factual information about the world. This implies that Twitter health news can be relied upon for insight into health services and behaviours.

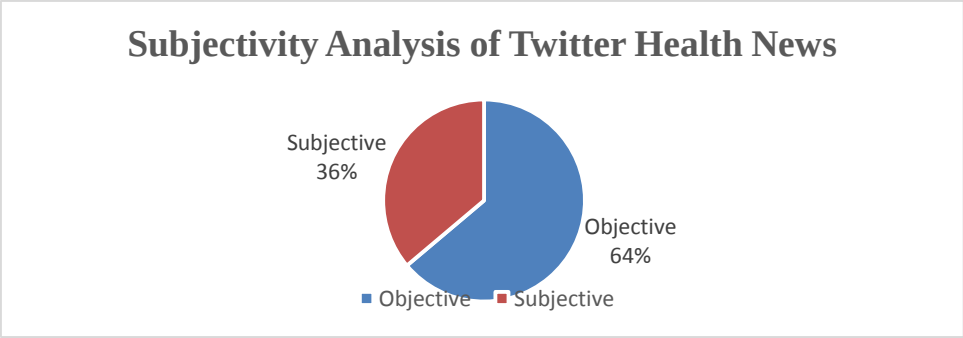

Fig. 2: Sentiment Subjectivity Analysis of Twitter Health News

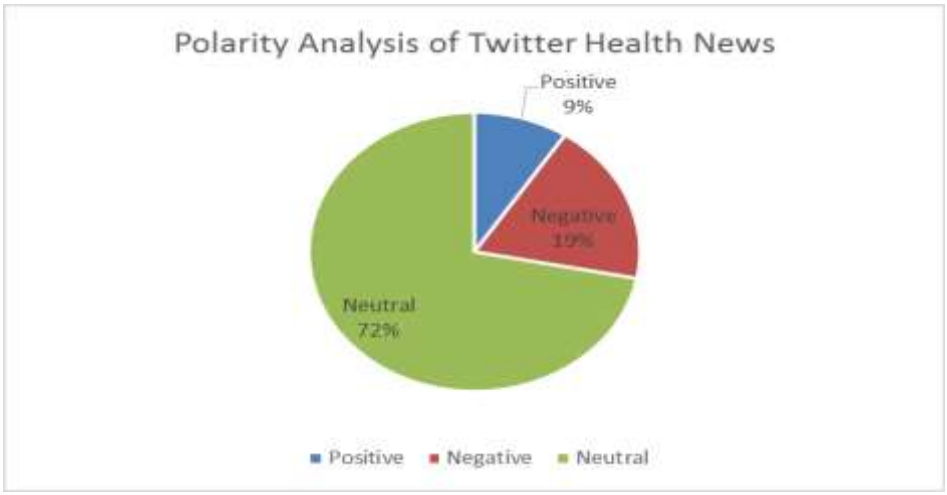

Fig. 3: Sentiment Polarity Analysis of Twitter Health News

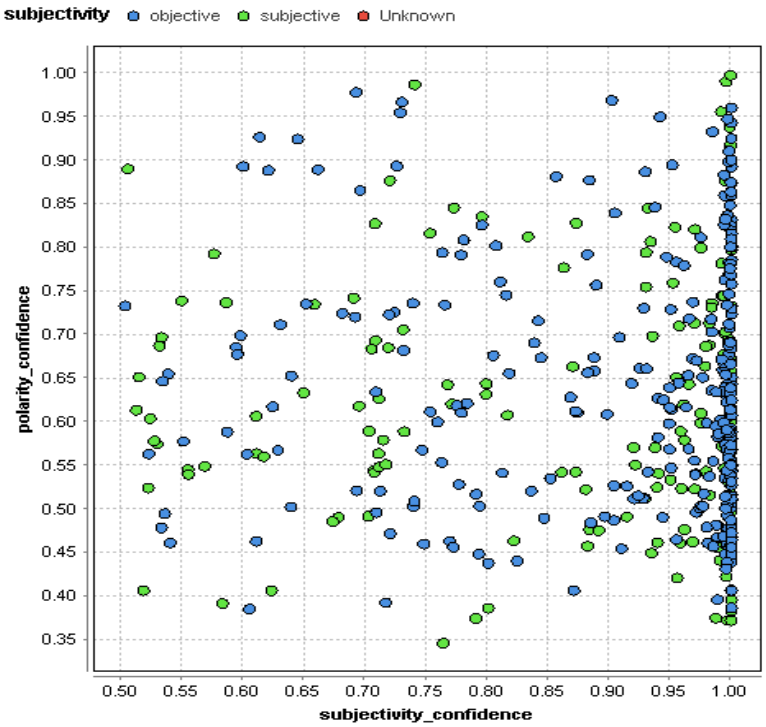

Fig. 4: Polarity/Subjectivity Confidence

Polarity confidence reveals the strength of sentiment polarity classification and it ranges from 0 to 1. A value close to 1 indicates higher confidence. From the analysis result (see figure 4), polarity confidence is ≈0.7 on the average.

The subjectivity confidence in the same vein reveals the strength of sentiment subjectivity classification which ranges from 0 to 1. A value close to 1 indicates higher confidence. From the analysis result, subjectivity confidence is 0.9 which indicates a high level of accuracy in the subjectivity classification (objective or subjective).

The polarity confidence (≈0.7) and subjectivity confidence (0.9) reveal that the result of the analysis is reliable.

CONCLUSION

This paper investigated general sentiments about Twitter health news. The focus is to determine whether users text in Twitter health news are expressing positive, negative or neutral opinion and whether the users are objective or subjective in their opinion. Natural language processing and text mining techniques (AYLIEN API) were used to extract sentiments subjectivities and polarities from a previously uncategorized dataset. A total of 24,588

tweets were analysed in all. The result shows that most of the tweets in Twitter health news are objective, that is, expressing factual information about the world with an average of 64% objective while 36% are personal views or opinion (subjective) with subjectivity analysis confidence of 0.9. Sentiment polarity reveals 9% positive, 19% negative and 72% neutral with polarity analysis confidence of 0.6. The polarity confidence (≈0.7) and subjectivity confidence (0.9) reveals that the result of the analysis is reliable.

RECOMMENDATION FOR FUTURE RESEARCH

In this research work, text analysis tool, AYLIEN API was used for sentiment analysis of the Twitter health news. For further work, this analysis can be repeated with other sentiment analysis APIs such as IBM Watson Tone Analyzer API, Qemotion, and PreCeive API for comparative analysis.

REFERENCES

Al-Amrani, Y., Lazaar, M., and El-Kadiri, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science, 127*, 511-520.

Godea, A. K., Caragea, C., Bulgarov, F. A., and Ramisetty-Mikler, S. (2015). An analysis of Twitter data on e-cigarette sentiments and promotion. In J. Holmes, R. Bellazzi, L. Sacchi, and N. Peek (Ed.), *Artificial Intelligence in Medicine. 9105*, pp. 205-215. Cham: Springer. doi:https:/doi.org/10.1007/978-3-319-19551-3_27

Greaves, F., Ramirez-Cano, D., Millet, C., Darzi, A., and Donaldson, L. (2013). Use of sentiment analysis for capturing patient's experience from free-text comments posted online. *J. of Med. Int. Res., 15*(11). doi:http://dx.doi.org/10.2196/jmir.2721

Hassan, A., Moin, S., Karim, A., and Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Application, 23*(11), 1-15.

Ji, X., Chun, S. A., and Geller, J. (2013). Monitoring public health concerns using twitter sentiment classification. *Proceedingsof the 2013 IEEE International Conference on Healthcare Informatics, ICHI'13* (pp. 335-344). Washington, DC, USA: IEEE Computer Society. doi:http://dx.doi.org/10.1109/ICHI.2013.47

Kallas, P. (2017). *Top 15 most popular social networking sites and apps.* Retrieved from DreamGrow: https://www.dreamgrow.com/top-15-most-popular-social-networking-sites

Katragadda, S., Benton, R., and Raghavan, V. (2017). Framework for real-time event detection using multiple social media sources. *Proceedings of the 50th Hawaii International Conference on System Sciences HICSS*, (pp. 1716-1725).

Korkontzelos, I., Nikfarjam, A., Shardlow, M., Sarker, A., Ananiadou, S., and Gonzalez, G. H. (2016). Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of Biomedical Informatics, 62*, 148-158.

Kranjc, J., Smailovic, J., Podpecan, V., Grcar, M., Znidarsic, M., and Lavrac, N. (2015). Active learning for sentimant analysis on data streams: Methodology and Workflow Implementation in the ClowdFlows platform. *Information Processing and Management, 51*, 187-203.

Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya, and F. J. Damerau, *Handbook of Natural Language Processing* (pp. 627-666). New York: CRC Press Tylor and Frncis Group.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*, 1-167.

Nasraoui, O. (2008). Book review: Web data mining - exploring hyperlinks, contents and usage data. *SIGKDD Explorations*, *10*(2), 23-25.

Ottoni, R., Las-Casas, D., Pesce, J., Meira Jr., W., Wilson, C., Mislove, A., and Almeida, V. (2014). Of pins an tweets: investigating how users behave across image-and text-based social networks. . *ICWSM*.

Panagiotou, N., Katakis , I., and Gunopulos, D. (2016). Detecting events in online social networks: definitions, trends and challenges. In S. Michaelis, N. Piatkowski, and M. Stolpe, *Solving Large Scale Learning Tasks: Challenges and Algorithms* (Vol. 9580, pp. 42-84). Cham: Springer. doi:https://doi.org/10.1007/978-3-319-41706-6_2

Pang, B., and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1-135. Retrieved from http://dx.doi.org/10.1561/150000001

Petz, G., Karpowicz, M., Furschub, H., Auinger, A., and Winkler, S. M. (2012). On text preprocessing for opinion mining outside of laboratory environments. *Active Media Technology*, 618-629.

Prieto, V. M., Matos, S., Alvarez, M., Cacheda, F., and Oliveira, J. L. (2014). Twitter: A good place to detect health conditions. *PloS, 9*(1).

Ribeiro, P. L., Weigang, L., and Li, T. (2015). A unified approach for domai-specific tweet sentiment analysis. *18th International Conference on Information Fusion* (pp. 846-853). IEEE.

Salas-Zarate, M. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodriguez-Garcia, M. A., and Valencia-Garcia, R. (2017). Sentiment analysis on tweets about diabetes: An aspect-level approach. *Computational and Mathematical Methods in Medicine*, 9 pages. doi:10.1155/2017/5140631

Schwartz, J., and Stone, A. (1998). Strategies for analysing ecological momentary assessment data. *Health Psychol., 17*(1), 6-16.

Statista. (2018, May 31). *Social media and user-generated content*. Retrieved from Statista: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Thakkar, H., and Patel, D. (2015). Approaches for sentiment analysis on twitter: A state-of-art study. *arXuv preprint arXiv:1512.01043*, 1-8.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervisd classification of reviews. *Procedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 417-424). Phildelphia, Pennsylvania: Association for Computational Linguistics.

Yoon, S., Elhadad, N., and Bakken, S. (2013). A practical approach for content mining of tweets. *American Journal of Preventive Medicine, 45*(1), 122-129. doi:10.1016/j.amepre.2013.02.025