Facebook User Account Data Analysis using NLP on A Specific dataset Selection Model

Nzenwata U.J
Department of Computer Science
Babcock University
Email: uchennajerry [AT] gmail.com

Adegbenle A.A
Department of Computer Science
Babcock University

Olugbohungbe R.R
Department of Computer Science
Babcock University

Abstract - **As one sits upon a social web tool like Facebook, virtually all the activities carried out is saved for the sole purpose of monitoring the trend of users' activities. The saved trends are often collated as big data for future analysis. These data are gathered as individual dataset which constitutes an ample size of global social web datasets, often used in machine learning techniques. In this study NLP Specific dataset Selection Model was developed, which uses natural language processing tools to analyze the personal data ever produced by a specific Facebook account from the point of account creation to the point the dataset was extracted was developed. This helps to gain insights about the vocabulary changes of a Facebook user over time. The result presented basic statistics on frequency of word usage and also present the shift in vocabulary usage overtime.**

Keywords: - ***NLP, Machine Learning, Lemmatization, Word Stemmer, JSON.***

I. INTRODUCTION

There is the individual thoughts on the amount of text data that is generated in a day over the social web spaces like the facebook, Linkedln, Twitter etc. A social web space user rarely concerns about the amount of text data exhumed off while sitting in face of the computer systems. The acts of being engaged on the social links are immeasurable to the extent that users rarely focus on the amount of text used in communication over a specific periods of time. These data are gathered as individual dataset which constitutes an ample size of global social web datasets, often used in machine learning techniques.

Social web data constitutes one of the major unstructured datasets [1]. An unstructured dataset is a form of big data, which is noted to be large amounts of unlabeled data. These sorts of data may have an internal structure, they are still considered "unstructured" because the data they contain doesn't fit neatly in a database [2].

There exist numerous social websites of which the Facebook appears to be the most prolific amongst all others. According to the Statistics and Studies from more than 22,500 Sources, Facebook had recorded 2.27 billion monthly active users as of the third quarter of 2018. Also, within the same period, 95.1 percent of active user accounts accessed the Facebook social network via smartphone. Hence, the Facebook platform is the most popular social network worldwide [3].-

The sole aim of this study is to develop a model that uses natural language processing tools to analyze the personal data ever produced by a specific Facebook account from the point of account creation to the point the dataset was extracted so as to gain insights about the vocabulary changes of Facebook users over time.

The remainder of this study is organized as follows section 2 provides literature review on related works, Natural Language processing, JSON Documents, and NLTK. Section 3 emphasizes on the explanation of the analytical methodology used. Section 4 presents the outcome of the analysis. Section 5 concludes the study with recommendations.

II. LITERATURE REVIEW

[4], present Sentiment Analyzer (SA) that extracts sentiment (or opinion) about a subject from online text documents. SA detects all references to the given subject, and determines sentiment in each of the references using natural language processing (NLP) techniques. The sentiment analysis consists of a topic specific feature term extraction; sentiment extraction, and (subject, sentiment) association by relationship analysis. SA utilizes two linguistic resources for the analysis: the sentiment lexicon and the sentiment pattern database. The performance of the algorithms was verified on online product review articles ("digital camera" and "music" reviews), and more general documents including general webpages and news articles. But this study could not classify the sentiment of an entire document about a subject.

NLP techniques have been used in Data extraction and data analysis from a body of document. In [5] for instance, the authors developed a text summarization system to enhance productivity and reduce errors in the traditional data extraction process. Using a system that utilizes machine learning and NLP the proposed approach can automatically generate abstracts of full-text scientific studies. The computer-generated abstracts were reported as a good source of information for data extraction while conducting a secondary study.

A study conducted by [6] provides an overview and tutorial of natural language processing (NLP) and modern NLP-system design. The description of the historical evolution of NLP, and the summary of common NLP sub-problems in this extensive field were elucidated. The study also made a brief description of common machine-learning approaches that are being used for diverse NLP sub-problems. The discussion on how modern NLP architectures are designed, with a summary of the Apache Foundation's Unstructured Information Management Architecture was found appreciated in thus study. Finally, the study considered possible future directions for NLP in which data processing was spear-headed.

The study of [7] reviews and discusses the use of natural language processing (NLP) and machine-learning algorithms to extract information from systematic literature. The study made progress in developing algorithms for automated annotation of taxonomic text, identification of taxonomic names in text, and extraction of morphological character information from taxonomic descriptions. Also, one of the significant studies [8], which is the first to demonstrate the use of high-level NLP techniques for qualitative data analysis presented a case study of the use of NLP for qualitative analysis in which the NLP rules showed good performance on a number of codes. In [9], the potential of using natural language processing to systematize analysis of qualitative data, and to inform quick decision-making in the development context was explored.

2.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a tract of Artificial Intelligence and Linguistics, devoted to make computers understand the statements or words written in human languages. Natural language processing came into existence to ease the user's work and to satisfy the wish to communicate with the computer in natural language. Since all the users may not be well-versed in machine specific language, NLP caters those users who do not have enough time to learn new languages or get perfection in it [10].

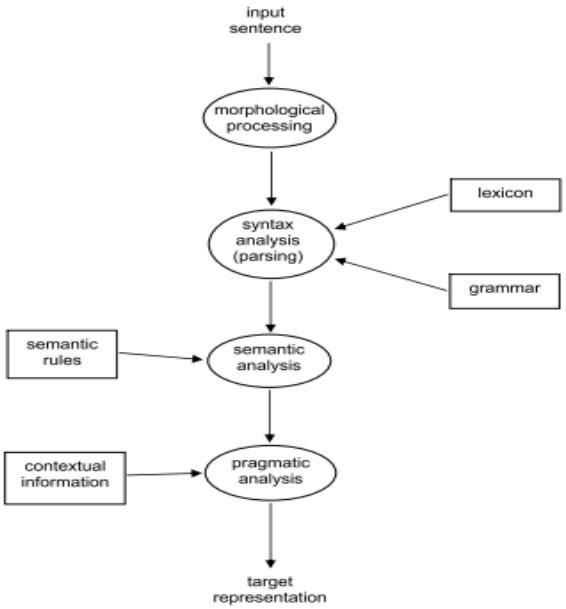

Figure 1: Broad Classification of NLP [10].

Linguistics is the science of language which includes Phonology that refers to sound, Morphology word formation, Syntax sentence structure, Semantics syntax and Pragmatics which refers to understanding [10].

Figure 2: The logical steps in Natural language Processing [11].

A simplified view of Natural Language Processing emphasizes four distinct stages (Figure 2). In real systems these stages rarely all occur as separated, sequential processes [11].

2.2 JSON Document Overview

JSON or JavaScript Object Notation is a lightweight text-based open standard designed for human-readable data interchange. Conventions used by JSON are known to programmers, which include C, C++, Java, Python, Perl, etc.

i. JSON stands for JavaScript Object Notation.
ii. The format was specified by Douglas Crockford.
iii. It was designed for human-readable data interchange.
iv. It has been extended from the JavaScript scripting language.
v. The filename extension is .json.
vi. JSON Internet Media type is application/json.
vii. The Uniform Type Identifier is public.json [12].

2.3 Natural Language Tool Kit (NLTK)

The Natural Language Toolkit (**NLTK**) is a platform **used** for building Python programs that work with human language data for applying in statistical natural language processing

(NLP). It contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning. It is a suite of open source Python modules with data sets and tutorials supporting research and development in natural language processing [13]. Its components are;

a. **Code**: corpus readers, tokenizers, stemmers, taggers, chunkers, parsers, wordnet, (50k lines of code)
b. **Corpora**: >30 annotated data sets widely used in natural language processing (>300Mb data)
c. **Documentation**: A 400-page book, articles, reviews, API documentation.

III. METHODOLOGY

This section highlights thoroughly the processes involved in achieving the specific aim of this study.

3.1 Data Source

The data source is fetched from the personal Facebook account. The dataset of interest here is the Messages (chats), Comments and the Posts. These dataset are in a JSON (Javascript Object Notation) format. Each dataset if differentiated by default using specified folders. The comments folder contains only one JSON file. The Posts folder contains two JSON files; your posts and other people's posts. The message folder contains several folders with unique ID values. Each folder in the message fold contains one JSON file.

3.2 Specific Data Selection

A module is developed which makes it easy to differentiate Facebook default dataset and the personal account dataset. This helps to dump the desired content into text files. Data is read in JSON. For each data folder that is available, the module checks if it is a message sent by default of by the personal account. The module also time stamps the message, and then add it to a file that takes the format of year.month.day.txt which is the format that enables any document update in vocabulary over time.

3.4 Load Data

Data Loading is the parsing a well selected dataset into the module-2 for data preprocessing.

3.3 Data Pre-processing

This technique is employed after data loading is done. A module is developed using NLTK tools. Here, the major work of data preprocessing is done and the output data in presented in four different ways. The nature of the output data are; Raw-data, Data with stopwords removed, stemmed data, and lemmatized data.

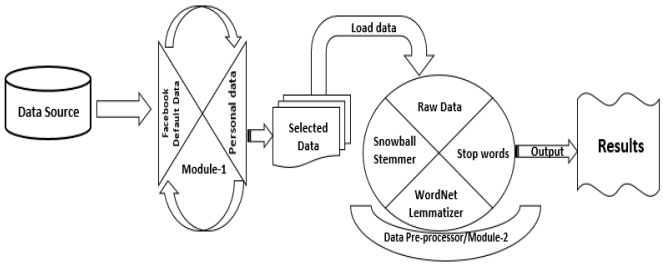

Figure 3: NLP on Specific dataset Selection Model

IV. RESULT PRESENTATION

The presented results are fragmented into desired output forms. These forms are the raw data form, Stemmer form, Lemmatized form, and stop words; presenting values of personal top words in values specified.

The result also generated some basic statistics on the personal data from inception of account creation as seen inn figure 2 bellow. The values of concern here are the top word, vocab size, average word usage count, Total words, and total characters from the account inception. We could also see that this personal Facebook account has 1,186 days of text activity on Facebook; and have used 121,316 characters so far.

Collection size:	1186
Top word:	u (1300)
Least common:	emoskihappy (1)
Vocab size:	3760
Average word usage count:	5.448404255319149
Total words:	32114
Total characters:	121316

Figure 2: Basis stats of the personal data.

The result also was able to present the shift in vocabulary usage overtime as evident in figure 3. Early 2015, it was deduced that the personal account user consumed a lot of vocabularies which amounts to excess time spent on Facebook making and sharing posts. Thus, the user was at the peak of Facebook activities within this period of time.

At mid-2015, it was obvious that there was declination in the user's activities on Facebook which may be attributed to busy schedules or loss of Facebook interest. See figure 4.

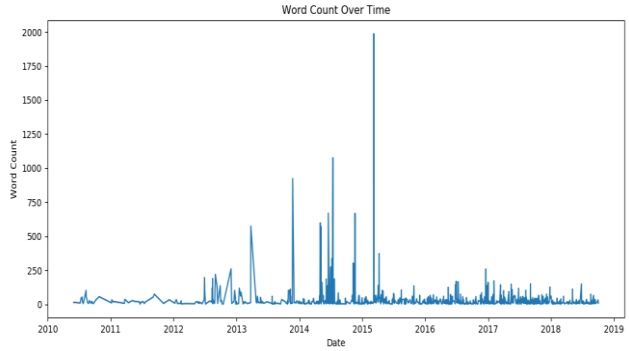

Figure 4: Vocabulary usage over time.

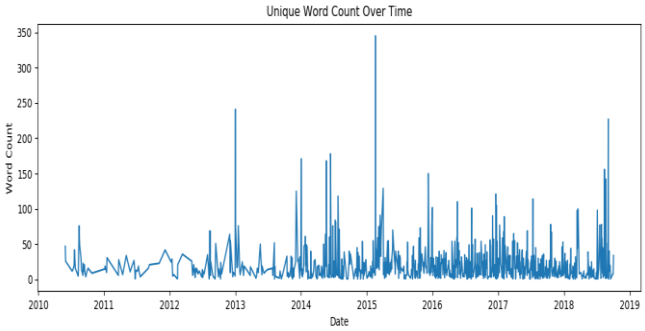

Figure 5: Unique word count over time

V. RECOMMENDATION AND CONCLUSION

This study did not incorporate Facebook API which would have made it scalable for organizations whose dealings are Facebook driven. This can be further developed to monitor the progress of the organizational outreach strength using the Facebook Analysis tools.

This study had shown the behavioral trend and changing observance of a Facebook user over a period of time. From the result output, users can determine the total word usage and the the frequency of each word used by the users. Also, Facebook users have the ability to monitor the vocabulary growth overtime the account remains active.

REFERENCES

[1]Wu, X., Zhu, X., Wu, G. Q., and Ding, W. (2014). Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), 97-107.

[2]Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 2053951716631130.

[3] "The Statistics Portal". (2018). Number of monthly active Facebook users worldwide as of 3rd quarter 2018 (in billions)

[4] Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. *Third IEEE International Conference on Data Mining*, 427–434. https://doi.org/10.1109/ICDM.2003.1250949

[5] D. D. A. Bui, G. Del Fiol, J. F. Hurdle, and S. Jonnalagadda. (2016). Extractive text summarization system to aid data extraction from full text in systematic review development. Journal of Biomedical Informatics, 64:265–272, 2016.

[6] Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

[7] Thessen, A. E., Cui, H., & Mozzherin, D. (2012). Applications of natural language processing in biodiversity science. *Advances in Bioinformatics*, *2012*(May 2014).

[8] Crowston, K., Allen, E. E., & Heckman, R. (2012). Using natural language processing technology for qualitative data analysis. *International Journal of Social Research Methodology*, *15*(6), 523–543. https://doi.org/10.1080/13645579.2011.625764

[9] Broniecki, P., & Hanchar, A. (2018). Data innovation for international development: An overview of natural language processing for qualitative data analysis. *Conference Proceedings - 2017 International Conference on the Frontiers and Advances in Data Science, FADS 2017*, *2018–Janua*, 92–97. https://doi.org/10.1109/FADS.2017.8253201

[10] Khurana, D., Koli, A., Khatter, K., & Singh, S. (2017). Natural Language Processing: State of The Art, Current Trends and Challenges, (Figure 1). https://doi.org/10.1111/jsr.12371

[11] Eggebraaten, T., Stevens, R., & Will, E. (2014). Natural language processing ('NLP-overview'). *US Patent 8,639,495*, 1–19. Retrieved from http://www.google.com/patents/US8639495

[12] Wall, L., Extraction, P., Language, R., Os, M., Scripting, P., Shell, U., … Point, T. (2015). About the Tutorial Copyright & Disclaimer. *Tutorial Points Pvt Ltd*, 2. https://doi.org/10.1017/CBO9781107415324.004

[13] Steven B., Ewan K., & Edward L. (2015). An overview of the Natural Language Toolkit. https://doi.org/10.1155/2012/391574