University of Ibadan
Journal of Science and Logics in ICT Research

Classification of Social Media Users by Interests and Sentiments using Text Mining Techniques

[1]Ajayi A. A. and Adeyemo A. B.
Computer Science Department, University of Ibadan,
[1]bdnajayi@gmail.com, [2]sesanadeyemo@gmail.com

Abstract

Social media sites like Twitter, Facebook, YouTube etc. have emerged as powerful platforms of communication where people share all kinds of information about topics they are interested in such as their opinion on real world events, personal experiences, product reviews and many more. The problem with this information is that it is unorganised and unstructured, therefore, it is difficult to assess automatically and in bulk. Studies on Twitter data have demonstrated that aggregating millions of messages can provide valuable insights into the interests of a population and opinions about said population. This study is aimed at gaining insights from the ever-growing Nigerian data generated from twitter by profiling the user to determine their interests and opinions. A framework for topic extraction and opinion mining is developed. The study used datasets across 5 popular and verified users in Nigeria to evaluate the proposed framework for its reliability and validity. Topic modelling was used to extract the topics of interest to the user while sentiment analysis was used to detect their opinions in each of their tweets which was further aggregated over each topic to get their total sentiscore about each topic. Topics of interests and overall interest level were detected within the timeframe of the datasets for the users. The interest of the users was obtained and compared among the last 6 months under observation to determine how users' opinions and interests changes over time. The findings, therefore shows that even though opinions and interests do change over time, the changes are generally minimal in subsequent months.

Keywords: *Text Mining, Interest and Sentiment Mining, Classification algorithms*

1. INTRODUCTION

Social media sites like Twitter, Facebook, YouTube etc. have emerged as powerful platforms of communication where people share all kinds of information about topics they are interested in such as their opinion on real world events, personal experiences, product reviews and many more. A large-scale study of about 58,000 Facebook users performed by Kosinski et al. reveals that digital records of human activity can be used to accurately predict a range of personal attributes such as age, gender, sexual orientation, and political orientation [4]. The explosion of the Web 2.0 has not only brought a huge volume of opinionated data recorded in digital forms, but also provided a great opportunity to understand the interest and sentiment of the public by analysing these large-scale data. However, all of the user generated data is a double-edged sword: the larger the size of the data, the more difficult it is to extract useful information. A survey shows that Facebook generates 250 million posts per hour and Twitter users on the other hand generate 21 million tweets per hour.

As social media continues to grow, more users share information about diverse topics and minutiae details about their everyday experiences. Users' interests which are usually influenced by day-to-day experiences and interactions with other users, changes rapidly, often several times daily. Currently users are made to preselect topics or categories they are interested in which are then used to classify the user. These topics of interests are used to guide users, recommend friends, posts and targeted advertisement.

Syntactic and Semantic approaches to extracting topics of interest and opinions

Ajayi A.A., and Adeyemo A. B. (2021). Classification of Social Media Users by Interests and Sentiments using Text Mining Techniques. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)*, Vol. 6 No. 2, pp. 79 – 89.

towards them have been successful when applied to documents with well-structured sentences, but as tweets do not necessarily maintain grammatical rules, direct analysis of tweets is not usually possible. Also, due to character limits of tweets, tweets are composed of abbreviations, irregular expressions and street lingo which changes frequently so the knowledge sources also have to change just as frequently.

As social media continue to evolve, users' interests change over time, sometimes several times during a day. This research paper presents a technique to capture and properly identify current interests of users at any point in time, categorize these interests and get a proper reflection of not only what they are interested in, but also their sentiments to such topics, in order to properly classify the user and better tailor their experiences.

2. RELATED WORKS

Social profiling or classification is the process of constructing a user's profile using his or her social data. In general, profiling refers to the data science process of generating a person's profile with computerized algorithms and technology [2]. A user's profile can be a combination of a number of things, including but not limited to [9],

1) a user's manual selected interests,
2) user's search history,
3) personal social network data.

Content sharing and creating has been the primary online activity of general social media users and that has forever changed online marketing. In the book 'Advanced Social Media Marketing' [14], the author gives an example of how a New York wedding planner might identify his audience when marketing on Facebook. Some of these categories may include:

1) Who live in the United States;
2) Who live within 50 miles of New York;
3) Age 21 and older;
4) engaged female.

Profiles must also be classified according to the kind of subject they refer to. This subject can either be an individual or a group of people. A group profile can also refer to a section of people that do not form a community, but are found to share previously unknown patterns of behaviour or other characteristics [1]. In that situation the group profile describes specific characteristics of a category of people, for example men with grey hair and sport cars, or teenagers with mobile phones and internet access. When a profile is constructed with the data of a single person, this is called individual profiling. This kind of profiling is used to identify the peculiar characteristics of a specific individual, in order to enable unique identification or the provision of customized services and offers [2].

The two major categories that exist in individual profiling are demographic and Psychographic profiling. Demographic profiling is a form of analysis used by marketers in order to be as accurate and efficient as possible when advertising goods or services and distinguishing any potential gaps in their strategy. By targeting certain groups who are more likely to be interested, a company can more efficiently and effectively utilize advertising resources so that they may get the maximum sales volume possible [6]. A recent discovery that has radically changed the way demographic profiles are generated, is the use of metadata. Metadata is the digital footprint left behind of everyone who uses online services. The more far-reaching a user's usage, the more the information available about them and their interests. The collection of metadata has proven to be a controversial topic, with large sections of the world populace expressing unease at the thought of their personal information being used to generate a virtual profile of them for businesses to take advantage of [7].

Psychographics meanwhile is a qualitative methodology used to describe traits of humans on psychological attributes [12]. Psychographics have been applied to the numerous studies of disposition, principles, opinions, interests, and lifestyles of people. Because this area of research focuses on activities, interests and opinions. Psychographic features are sometimes abbreviated to 'AIO variables'. Psychographics are applied to the study of cognitive attributes such as attitudes, interests, opinions, and belief,

as well as the study of overt behaviour (e.g., activities). The opinions and reactions that are shared using social networking platforms appear to be the most effective way of deducing a user's personality. Psychographic methods gained importance during the 2016 US presidential election campaigns of both Hillary Clinton and Donald Trump, with the latter broadly using them in targeted advertisements to narrow down his constituencies.

Topic modelling is an often-used text-mining tool for discovering hidden semantic structures in a text body. It is an attempt to inject semantic meaning into vocabulary. Topic models are a useful and ubiquitous tool for understanding large corpora and is a useful mechanism for identifying and depicting various concepts embedded in a document collection allowing the user to be able to navigate the collection in a topic guided manner. Topics made up of meaningful words, provide the user with an extensive overview of the content of the document collection. Each document is represented as a combination of automatically constructed topics and the user may select documents related to a specific topic of interest and vice versa [5].

Opinion mining (also known as sentiment analysis) is the computational study of opinions, sentiments and emotions expressed in text [5]. The field of sentiment analysis and opinion mining is well-suited to various types of intelligence applications. Indeed, business intelligence seems to be one of the main factors behind corporate interest in the field [10]. Approaches for opinion mining can be broadly classified into machine learning-based method and lexicon-based method. Machine learning based approaches for opinion mining are supervised learning task. They utilize textual feature representations coupled with classification algorithms to infer the opinions expressed in the text [13]. Unsupervised lexicon-based techniques rely on the assumption that the collective polarity of a sentence is the sum of polarities of the individual words or phrases of that sentence [3]. An unsupervised lexicon-based system was proposed by [11], they used Word Net to classify the text using an assumption that texts with similar polarity have similar orientations.

3. MATERIALS AND METHODS

Using the twitter data available, profiles were created for twitter users based on their previous tweets by determining topics that they are interested in within a given time and their opinions about those topics. The methodology will consist of 5 steps or stages (Figure 1) namely:

i. data collection,
ii. data pre-processing using Tokenization and Bag of Words representation,
iii. topic modelling using an unsupervised model,
iv. sentiment analysis using semantic orientation approach,
v. evaluation

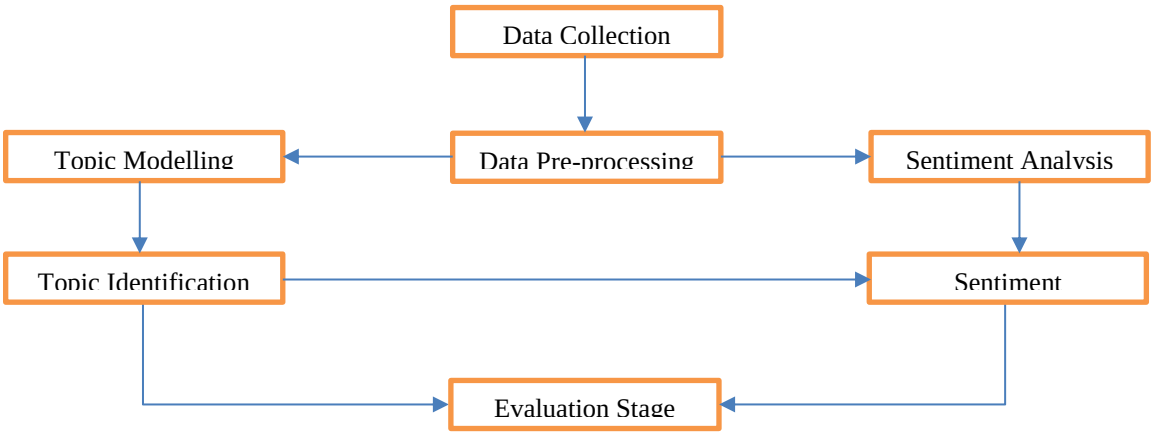

Figure 1: Methodology

3.1 Data Collection

Tweets are messages posted on Twitter social networking platform by users of the platform to convey their thoughts and opinions about certain events, people, item or activity to their friends and the world at large.

Five most popular verified Nigerian Twitter accounts or users' handles were identified in order to obtain the data for this research and they are @MBuhari, @OgbeniDipo, @atiku, @davido and @seyiamakinde, and the aggregation of the dataset obtained is shown in Table 1.

To retrieve these tweets from these accounts, the Tweepy Python library was used. This library allows the retrieval of historical, as well as current tweets from the Twitter platform unlike the Twitter Streaming API which allows retrieval of tweets only a few weeks old.

3.2 Pre-processing

This is a crucial stage where the noise is drastically reduced from the tweets' dataset. Tweets contains a lot of data which are unnecessary for our evaluation purposes such as slangs, abbreviations, hashtags, URLs, email addresses, punctuations etc. Therefore, there is need to strip the data of these unnecessary content and it involves the following steps:

1. Duplicates Removal: It is common for people to copy and paste quotes and retweet post which they might previously done. It is therefore imperative for such duplicates to be removed.
2. Normalization: As there is no grammatical rule enforced about how text in a tweet should be composed, tweets can contain words with different cases. In order to ensure that same words written in different cases will not be considered as being different, Normalization, a method which converts all tweets into lowercase is performed.
3. Remove username mentions: Twitter enables including usernames within tweets through the symbol "@". Usernames of users that the original user intends to be notified or targeted of the tweet are added to the tweet and are also referred to as "mentions". These do not possess any value for the analysis, hence they need to be removed from the dataset using a function.
4. Remove retweet tags and hashtags: A retweet is a repost or quote of a tweet by another Twitter user. This is done either to make comments about the original tweet, indicate support or approval of the original tweet or simply just pass it along as the user's own opinion. In any of these circumstances the retweeted post is taken as the user's opinion and so the retweet tag "RT", which is there to show that it is a retweet is not necessary for this evaluation and should be removed. Same as with retweet tags, hashtags also are considered not of significant value for topic modelling analysis, in particular and therefore are removed.
5. Remove URLs: URLs are not words with meanings and hence will not be useful for this evaluation. In other to remove URLs from the dataset, regular expression was used to define and identify the patterns of URLs and replace them with an empty space.
6. Digits removal: Only words are of interest in this evaluation. So, using regular expressions every occurrence of digits with is replaced with an empty space.
7. Remove punctuations, double spacing and new line characters: Double spacing could will be created by removal of all the above. For this evaluation these constitute unnecessary noise and should be removed as they will make our data harder to process while also not letting us get valid results.
8. Short words removal: Words with less than three characters (short words) are removed from the dataset, simplifying feature extraction for the analysis.
9. Tokenization: An essential step of pre-processing is known as Tokenization. It is the process where the text is split according to whitespaces, and every word and punctuation is saved as a separate token.
10. Stopword removal: Stopwords are words that occur often in texts and most times do not convey the core idea of the text. These are referred to as stopwords. They have no analytic value and are usually articles, prepositions, or pronouns, for instance, 'a,' 'and,' 'the,' etc.
11. Drop Tweets with Less than 3 Tokens: Tweets with less than three tokens are removed, this results in less documents to be considered further on in the analysis phase. For topic modeling and also sentiment

analysis, documents with less than three tokens are not suitable to generate enough information.

12. Lemmatization: A word is sometimes used in different forms of its root form. For instance, the word sing can be used as sang, sung, singing all of which come from the same root word sing. Lemmatization is the process of deriving the root form of words by removing prefixes and suffixes.

3.3 Topic Modelling

Topic modelling is an unsupervised machine learning algorithm for discovering topics in a collection of documents. In this case the collection of documents is a collection of tweets. Topic models are suitable for unlabelled data and they give us insights into the corpus [17], letting us know more about the corpus in more specific ways beyond a general term like health, politics or sports.

As stated earlier, topic modelling is the process of applying statistical models (topic models) to extract the hidden topics from a dataset. To perform topic modelling on our dataset the following steps need to be performed:

1. Bigram and Trigram Generation
2. Lemmatization and Stemming
3. Dictionary and Corpus Generation
4. Topic Model Generation

3.4 Bigram and Trigram Generation

Bigrams and Trigrams are a sequence of words which often occurs together when expressing a particular meaning. A sequence of N words is known as N-Grams, as theoretically N can be of any length; the most common are pairs of words (Bi-grams) like “please turn”, “turn your”, or “your homework” and a sequence of three words (Tri-grams) like “please turn your”, or “turn your homework”.

The bigram and trigram model approximate the probability of a word given all the previous words by using only the conditional probability of one preceding word. In other words, it is approximated with the probability: P(w | h). P(w | h) the probability of word w, given some history, h. And so, when a bigram model is used to predict the conditional probability of the next word, the following approximation are made:

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-1}) \quad (1)$$

3.5 Lemmatization and Stemming

The next step is lemmatization, an essential step for many applications of text mining. Lemmatization takes into consideration the context and converts the word to its base form; for instance, the term “hugging” is converted to “hug” and “best” to “good.” For the lemmatization task, the package used is spaCy, an open-source library with many pre-built models for natural language processing.

After this step, duplicates will be dropped once more, as short tweets composed of very few tokens when lemmatized may lead to duplicate rows.

Another useful technique is word stemming, which is the process of transforming a word into its root form. Unlike lemmatization, stemming is a more aggressive approach as suffix cuttings often result in non-meaningful English words. For instance, the word “animals” would be lemmatized as “animal,” but the Porter Stemmer yields “anim.” It was decided to implement both to help in dimensionality reduction.

3.6 Dictionary and Corpus Generation

Before building the LDA model, two main inputs: the dictionary and the corpus, which are created using functions from the Gensim package, have to be created. After stemming the data, it needs to be represented in a form that can be fed into the topic modelling algorithm. To do this, all unique words in the corpus are first generated and each word assigned a unique token.

The algorithm interacts with this unique token generated as a representation of the words in the corpus. The dictionary module in Python Gensim library is used to generate the dictionary of the corpus.

Next, the corpus is represented in terms of each document and the unique words in the corpus.

Each document also has a unique index. Finally, a matrix is generated in which column represents a term while each row represents a document and the entries of the matrix represents the frequencies of the terms in the documents. These entries could be raw counts, TF or TF-IDF. To get the bag-of-words representation of the corpus, the doc2bow function of the dictionary module in Python Gensim library is used.

Gensim assigns a unique Id to each word, and then the corpus is represented as a tuple (word_id, word_frequency).

3.7 Topic Model Generation

For the mining of the data, topic modelling has been chosen to be used as they are unsupervised algorithms which are suitable for unlabeled data as twitter data. Secondly, topic models give insights into a corpus, letting us know what the corpus is about in more specific ways beyond a general term like health or sports.

As earlier stated, topic modelling is the process of applying statistical models (topic models) to extract the hidden topics in the data. However, there are several topic modelling techniques available and Latent Dirichlet Allocation (LDA) has been chosen to be used as it has been recurrently recommended as being good for text mining from literature [17].

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus [16]. This means that LDA describes how the corpus is generated using a probabilistic model which is precisely the Dirichlet distribution. The Dirichlet distribution is a probabilistic distribution over other probability distribution. Dirichlet best describes the generation of a corpus form various topics because naturally all topics will not follow the same probability distribution in the generation of the corpus. With LDA, documents are represented as random mixtures over latent topics and each topic is describes by a distribution over words [16].

The generative process assumed by LDA is as follows:

1. Choose N ~ Poisson(ξ).
2. Choose θ ~ Dir(α).
3. For each of the N words *wn*:
   a. Choose a topic zn ~ Multinomial(θ).
   b. Choose a word wn from (wn |zn, β) a multinomial probability conditioned on the topic zn.

Hence, given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \quad (2)$$

3.8 Sentiment Analysis

The main task of sentiment analysis is sentiment classification, which is classifying the polarity of the text. Most of the sentiment analysis researches based on the semantic orientation (SO) approach have mainly investigated document-level or sentence-level classification [15]. For this stage the TextBlob python Library was used.

TextBlob is a Python library for processing textual data. It provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more. The sentiment property returns a named tuple of the form Sentiment (polarity, subjectivity). The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.

The pre-processing phase is slightly different from the one employed for Topic Modeling because of the features that TextBlob sentiment analyzer embodies, the steps taken are similar to the same steps for the pre-processing phase for topic modeling with the addition of keeping of stopwords as the TextBlob sentiment analyzer takes them into consideration.

4. RESULTS AND DISCUSSION

4.1 Results

After the topic modelling and sentiment analysis was performed, the results were grouped per user over each topic of interest and a total_sentiweight, which is the total addition of the sentiweight of all the tweets under that topic

over the period of time under consideration is obtained.

The total_sentiweight reflects the level of interest that the user has for the particular topic over the period under consideration, while considering the user's sentiment about each tweet that falls under the topic.

Running the framework on a user's tweet dataset yielded the following topic keywords, topic label and total_sentiscore for the user as shown in table 2 and mapped in Figure 2.

The user was interested in the topic "Peaceful Nation" the most with 1181 tweets and a total_sentiscore of 361.0, followed by "Governance Continuation" with 531 tweets and a total_sentiscore of 133.0. The user was the least interested in the topics "News" and "Election" with total_sentiscore of 10.0 and total_sentiscore of 23.0 respectively.

The framework was run over 6 months of the dataset in order to be able to get more insight into the interests of the user over that period of time and the result is shown in Table 3. Also comparing the results obtained over this period will enable analysis of how the interest of the user changes over time as shown in Figure 3.

Table 1: Tweet Database Analysis

User Handle	**Number of Tweets**	**Number of Non-Duplicate Tweets**	**Number of Cleaned Tweets**
@MBuhari	3219	3213	3163
@OgbeniDipo	3123	3072	2529
@atiku	3189	3189	3061
@davido	3094	3079	2079
@seyiamakinde	3223	3215	3073
Total Number	**15848**	**15768**	**13905**

Table 2: Topics keywords and total sentiweight

ID	**topic**	**Topic Keywords**	**Tweets**	**Total_SentiWeight**
0	Election	0.089*"fair" + 0.047*"presid" + 0.045*"time" + 0.038*"build" + 0.034*"return" + 0.025*"even" + 0.024*"victori" + 0.022*"complet" + 0.021*"call" + 0.017*"say"	57	23.0
1	Infrastructure	0.102*"year" + 0.078*"state" + 0.076*"last" + 0.040*"visit" + 0.034*"effort" + 0.030*"famili" + 0.024*"fight" + 0.022*"infrastructur" + 0.020*"home" + 0.019*"attack"	186	66.0
2	African Vision	0.094*"tell" + 0.057*"take" + 0.037*"part" + 0.035*"leader" + 0.027*"join" + 0.023*"prayer" + 0.022*"speak" + 0.021*"african" + 0.017*"affect" + 0.014*"vision"	91	31.0
3	Peaceful Nation	0.147*"today" + 0.033*"report" + 0.032*"peopl" + 0.031*"host" + 0.029*"earlier" + 0.028*"audienc" + 0.027*"nigerian" + 0.026*"receiv" + 0.023*"peac" + 0.018*"nation"	1181	361.0
4	Administration	0.027*"administr" + 0.026*"never" + 0.025*"follow" + 0.024*"terror" + 0.023*"give" + 0.023*"alway" + 0.023*"recoveri" + 0.022*"kill" + 0.022*"deliv" + 0.021*"address"	111	35.0
5	Military Budget	0.090*"attend" + 0.044*"wish" + 0.034*"commun" + 0.031*"project" + 0.030*"militari" + 0.029*"sign" + 0.027*"afternoon" + 0.026*"discuss" + 0.022*"safe" + 0.020*"budget"	101	20.0

6	Economy	0.074*"work" + 0.073*"busi" + 0.033*"ensur" + 0.033*"meet" + 0.028*"power" + 0.026*"issu" + 0.024*"economi" + 0.020*"world" + 0.019*"invest" + 0.018*"plan"	300	95.0
7	Security Agencies	0.051*"offici" + 0.050*"privat" + 0.049*"secur" + 0.047*"expens" + 0.039*"must" + 0.029*"countri" + 0.024*"remain" + 0.023*"commit" + 0.022*"agenc" + 0.019*"well"	373	111.0
8	News	0.086*"mark" + 0.081*"wild" + 0.036*"success" + 0.036*"welcom" + 0.023*"news" + 0.022*"respect" + 0.020*"proud" + 0.019*"achiev" + 0.019*"governor" + 0.017*"forc"	34	10.0
9	Governance Continuation	0.075*"also" + 0.071*"make" + 0.068*"play" + 0.045*"continu" + 0.039*"govern" + 0.036*"possibl" + 0.036*"case" + 0.034*"clean" + 0.033*"gold" + 0.027*"support"	531	133.0
10	Approvals	0.054*"thank" + 0.034*"approv" + 0.032*"present" + 0.032*"releas" + 0.030*"presidenti" + 0.023*"member" + 0.021*"current" + 0.020*"next" + 0.020*"fulfil" + 0.017*"growth"	81	24.0

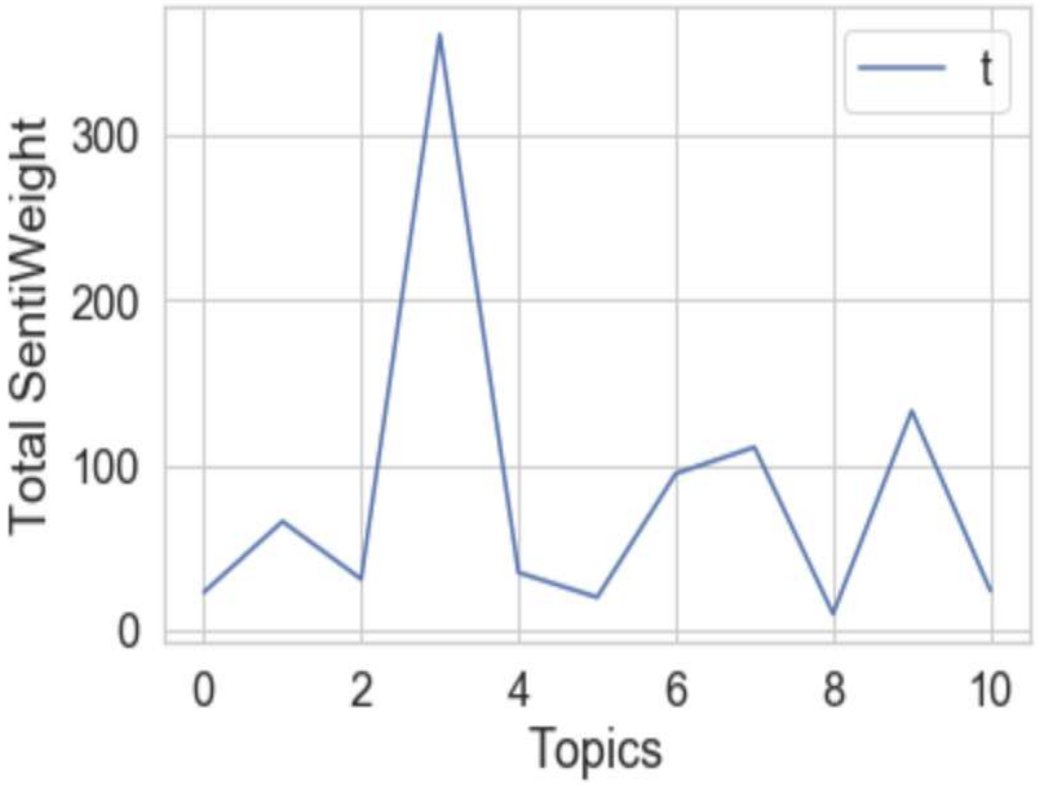

Figure 2: Total Sentiweight against Topics for the user

Table 3: Topic and Total Sentiweight over 6 months for the user

	Topic	Month 1 tweets	Month 1 total senti_ weight	Month 2 tweets	Month 2 total senti_ weight	Month 3 tweets	Month 3 total senti_ weight	Month 4 tweets	Month 4 total senti_ weight	Month 5 tweets	Month 5 total senti_ weight	Month 6 tweets	Month 6 total senti_ weight
0	Electi-on	4	1	1	1	3	1	2	0	1	1	3	2
1	Infrastr ucture	11	1	2	0	6	3	4	1	2	2	2	0
2	African Vision	9	2	0	0	9	0	0	0	2	0	1	1
3	Peace-ful Nation	46	15	6	1	51	11	10	3	11	4	5	1
4	Admini stration	5	1	1	0	4	0	2	1	3	1	2	-1
5	Militar-y Budget	2	0	2	0	4	1	0	0	1	-1	1	0
6	Econo-my	34	10	2	-1	18	2	4	1	4	-2	4	0
7	Securi-ty Agenci es	26	9	0	0	12	-1	2	2	6	1	5	2
8	News	5	3	0	0	4	3	0	0	0	0	1	1
9	Gover-nance Contin uation	34	9	3	1	34	8	1	0	2	1	7	2
10	Appro-vals	4	2	0	0	8	2	0	0	1	1	1	1

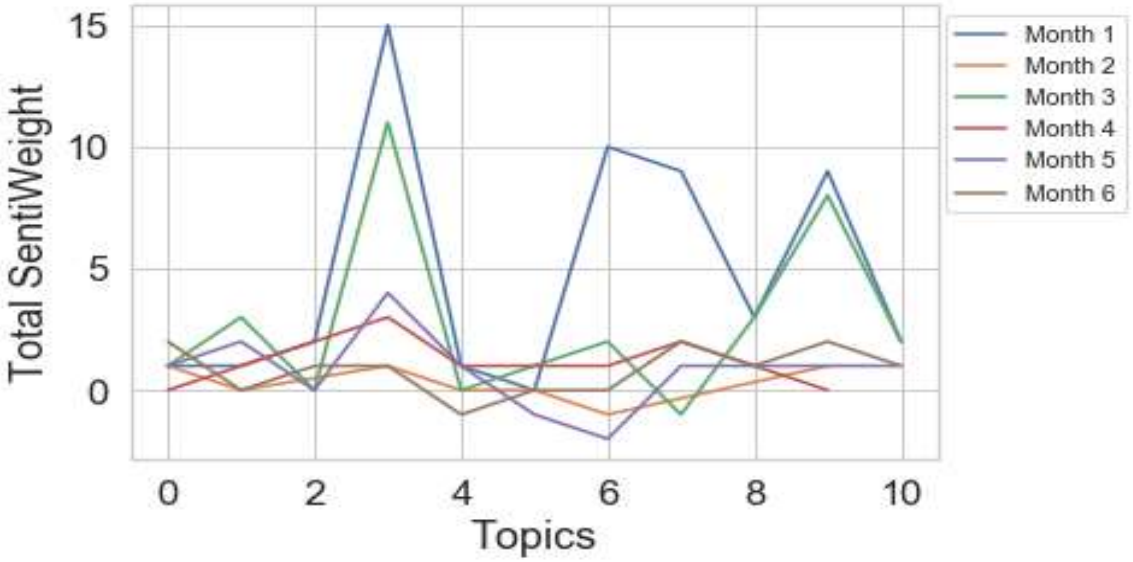

Figure 3: Topic and Total Sentiweight Comparison over 6 months for the user

The topics of the highest positive interest to the user are “Peaceful Nation” in the first month. The same topic was also of the highest positive interest in subsequent months except for the last month when the user’s highest interest was “Governance Continuation”. At the other end the user felt negative overall about a topic in the second, third, fifth and sixth month. The user felt negative about the topic “Economy” in the second month, “Security Agencies” in the third month, “Military Budget” and “Economy” in the fifth month and “Administration” in the sixth month.

A consistent pattern can be observed in the topic of most interest to the user either positively or negatively regardless of the number of topics of interest for that month which was 7 topics for the second and fourth months, 10 topics in the fifth month and 11 topics in the first, third and sixth months.

4.2 Discussion of Results

This research comes as a way to find an effective solution to be able to use opinionated data to profile a user automatically by determining both topics interested in and opinions about those topics. It extends and builds on previous works in the area of opinion mining, topic extraction and sentiment analysis previously referenced and extends the scope to enable capturing of opinions.

5. CONCLUSION

This study has developed a framework for finding an effective solution for using opinionated data to profile a user automatically by determining both topics interested in and opinions about those topics. The framework was evaluated using the prototype system implemented in the study.

References

[1] Custers, B. H. M. (2004). The power of knowledge . In *Ethical, legal, and technological aspects of data mining and group profiling in epidemiology* . Wolf Legal Publishers (WLP) .

[2] Kanoje, S., Mukhopadhyay, D., & Girase, S. (2016). User Profiling for University Recommender System Using Automatic Information Retrieval. *Physics Procedia*, *78*(December 2015), 5–12. https://doi.org/10.1016/j.procs.2016.02.002

[3] Kaushik, C., & Mishra, A. (2014). A Scalable, Lexicon Based Tchnique for Sentiment Analysis. *International Journal in Foundations of Computer Science & Technology*, *4*. https://doi.org/10.5121/ijfcst.2014.4504

[4] Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

[5] Liu, B. (2012). Sentiment Analysis and Opinion Mining. In *Synthesis Lectures on Human Language Technologies* (Vol. 5). https://doi.org/10.2200/S00416ED1V01Y201204HLT016

[6] Maureen, F., & David, A. (1996). Understanding demographic effects on marketing communications in services. *International Journal of Service Industry Management*, *7*(3), 31–45. https://doi.org/10.1108/09564239610122947

[7] Needel, S. P. (2020, 04 21). *Why Big Data is a Small Idea: And why you shouldn't worry so much*. Retrieved from Warc: https://www.warc.com/content/paywall/article/why_big_data_is_a_small_idea_and_why_you_shouldnt_worry_so_much/100226

[8] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

[9] Saoud, Z., & Kechid, S. (2016). Integrating social profile to improve the source selection and the result merging process in distributed information retrieval. *Information Sciences*, *336*,

115–128. https://doi.org/10.1016/j.ins.2015.12.012

[10] Shahheidari, S., Dong, H., & Daud, M. N. R. B. (2013). Twitter Sentiment Mining: A Multi Domain Analysis. *2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems*, 144–149. https://doi.org/10.1109/CISIS.2013.31

[11] Vanzo, A., Croce, D., & Basili, R. (2014). A context-based model for Sentiment Analysis in Twitter. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2345–2354. https://www.aclweb.org/anthology/C14-1221

[12] Wells, W. D. (1975). Psychographics: A Critical Review. *Journal of Marketing Research*, *12*(2), 196–213. https://doi.org/10.2307/3150443

[13] Zhou, X., Tao, X., Yong, J., & Yang, Z. (2013). Sentiment analysis on tweets for social events. *Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2013*, 557–562. https://doi.org/10.1109/CSCWD.2013.6581022

[14] Funk, T. (2013). *Advanced Social Media Marketing: How to Lead, Launch, and Manage a Successful Social Media Program*. Berkeley, CA: Apress.

[15] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

[16] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3 , 993–1022.

[17] Saif, H., He, Y., Fernandez, M., & Alani, H. (2014). Semantic Patterns for Sentiment Analysis of Twitter. *Proceedings of the 13th International Semantic Web Conference - Part II*. https://doi.org/10.1007/978-3-319-11915-1_21