Tracking Food Insecurity from Tweets Using Data Mining Techniques

Andrew Lukyamuzi
Mbarara University of Science and Technology, Faculty of Computing and Informatics, Uganda
andrewlukyamuzi@gmail.com

John Ngubiri
Makerere University, College of Computing and Information Sciences, Uganda
ngubiri@cit.ac.ug

Washington Okori
Uganda Technology and Management University, Uganda
gokori@gmail.com

ABSTRACT

Data mining algorithms can be applied to extract useful patterns from social media conversations to monitor disasters such as tsunami, earth quakes and nuclear power accidents. While food insecurity has persistently remained a world concern, its monitoring with this strategy has received limited attention. In attempt to address this concern, UN Global Pulse demonstrated that tweets reporting food prices from Indonesians can aid in predicting actual food price increase. For regions like Kenya and Uganda where use of tweets is considered low, this option can be problematic. Using Uganda as a case study, this study takes an alternative of using tweets from all over the world with mentions of; (1) uganda +food, (2) uganda + hunger, and (3) uganda + famine for years 2014, 2015 and 2016. The study however utilized tweets on food insecurity instead of tweets on food prices. In the first step, five data mining algorithms (D-tree, SVM, KNN, Neural Networks and N-Bayes) were trained to identify tweets conversations on food insecurity. Algorithmic performance were found comparable with human labeled tweet on the same subject. In step two, tweets reporting food insecurity were generated into trends. Comparing with trends from Uganda Bureau of Statistics, promising findings have been obtained with correlation coefficients of 0.56 and 0.37 for years 2015 and 2016 respectively. The study provides a strategy to generate information about food insecurity for stakeholders such as World Food Program in Uganda for mitigation action or further investigation depending on the situation. To improve performance, future work can; (1) aggregate tweets with other datasets, (2) ensemble algorithms, and (3) apply unexplored algorithms.

KEYWORDS

Food insecurity, Social Networks, Big Data, Web, Text Mining, Machine Learning Algorithms

ACM Reference Format:
Andrew Lukyamuzi, John Ngubiri, and Washington Okori. 2018. Tracking Food Insecurity from Tweets Using Data Mining Techniques. In *SEIA '18: SEIA '18: Symposium on Software Engineering in Africa , May 27–28, 2018, Gothenburg, Sweden.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3195528.3195531

SEIA '18, May 27–28, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5719-7/18/05…$15.00
https://doi.org/10.1145/3195528.3195531

1 INTRODUCTION

Social network technologies that have emerged such as Twitter, Flicker, Myspace, and Facebook have brought a revolution in connecting with one another. They are providing virtual environment for a large number of users to share ideas among themselves. Time and geographical location are no longer a big limitation. Twitter and Facebook are at the forefront in the arena of social networking. Twitter has more than 200 million users [1] while Facebook has more than 1.35 billion users [8]. Such a big number of users have culminated into enormous generation of communication data in these social networks. Facebook receives about 41,000 postings per second and Twitter about 6,000 tweets per second [37]. The users in these social networks are people from various walks of life and the conversations they generate also span various topical areas including but not limited to politics, religion, social affairs, economics, disasters and education. Phillips, et al. [23], assert that social network data provides a vast record of humanity's everyday thoughts, feelings, and actions at a resolution previously unimaginable. Unlike traditional data (for example survey data), social network conversations are less expensive to acquire, voluminous, generated at fast rate in real time and originate from big geographical area. This offers opportunities to study phenomena such as live situations, spatial patterns and temporal patterns. Data such as this is termed as big data and unique peculiarities of big data are the 3Vs: Variety, Volume and Velocity. Big data suffers from drawbacks such as being; noisy, unstructured and can be incomplete. The web is a gigantic environment hosting big data. By the end of 2016, Internet data was growing at rate of 1.1 billion terabytes per year and was projected to reach 2 billion terabytes per year in 2019 [14].

For a long period the societies have been dealing with traditional data sets. Stretching previous techniques for manipulation of traditional datasets on to big data has not guaranteed a happy path due to differences between these two data sets. For this has opened a new window to seek for more promising techniques to mine meaningful patterns from big data. The science of mining big data has led to new terminologies but all relating to data mining. Such terminologies include; data analytics, data science, machine learning and computational intelligence. When it comes to extracting patterns from Internet data the term big data mining or data mining transforms into web mining. Web mining falls into three categories, content, structure and behavior. Structure analysis examines linkages among sites, behavior analysis focuses on examining users actions on the web while content analysis deals with extraction of patterns from the contents of the web data. Now this research is on mining social network conversations (a web content type) to track food insecurity. Next we put the study into a specific perspective.

While several studies have emerged for mining social network to study disasters such as earth quakes, floods and tsunami, limited attention has been given on how the same strategy can be harnessed to track food insecurity. This study explorers this possibility. It is based on a premise that perturbations in food availability influence people's conversations towards food availability. This information can be shared in social networks and this is an opportunity to examine the current situation of food availability. The study closely relates to an investigation by UN Global Pulse [25] which used tweets from citizens to predict food price increase but for this study the focus is on tracking food insecurity. For regions like Kenya and Uganda where use of tweets is considered low, the option of using tweets from citizens can be problematic [30]. The study takes an alternative of using tweets from all over the world using Uganda as a case study. Uganda is in East Africa with 40.6 million people [33]. Most of its households derive their livelihoods from agriculture [32]. In past Uganda has experienced incidences of food insecurity for example in; 1980[36], 1997-1998Bahiigwa [3], and 2006[10]. Recently (in 2017) about 10.9 million people in Uganda were reported as victims of food insecurity [13]. Consequently, Uganda was a fair candidate case for this investigation. The tweets from all over the world were retrieved based on three sets of keywords; (1)uganda food, (2) uganda hunger, and (3) uganda famine for years 2014, 2015 and 2016. To generate the trends to track food insecurity, the study required only tweets holding conversations on food insecurity yet some tweets in the collection were not. To address this, five classifiers were trained to label relevant tweets (tweets holding conversations on food insecurity) from the irrelevant. Now the relevant tweets were generated into food insecurity trends. To establish validity of these results, the trends were compared with other trends generated using official alternative datasets. Trends from this study can reveal an emerging situation of food insecurity to alert stakeholders such as World Food Program- Uganda for appropriate action like preparing for mitigation.

The study was guided by four research questions: (1) How can tweets be used track food insecurity using data mining techniques? (2) How does the research relate to previous work? (3) How can data mining techniques be used to filter (label) the tweets required for the trends? (4) How can the generated trends be validated?

This paper has been organized into six sections. The next section is on related research identifying key issues the study addresses. Section 3 is a methodology, the means used to conduct the study. Section 4 presents results while section 5 discusses the results. The paper ends with section 6 which gives conclusion and future work.

2 LITERATURE REVIEW

Social networks are playing an instrumental role in tracking disaster patterns. In this section we highlight areas linked to this. Strengths and drawbacks of the current state of practices are examined. A link between the previous works with our proposed solution is studied to highlight which areas are adopted and where improvement has been brought forth.

Food insecurity is a key word in this study requiring contextualization. To appreciate the term food insecurity requires understanding first what food security is. Food security is "situation that exists when all people, at all times, have physical, social, and economic access to sufficient, safe, and nutritious food that meets their dietary needs and food preferences for an active and healthy life [39]."Food security rests on four dimensions; availability, stability, utilization and accessibility [9][24]. Availability is the capacity to meet the overall demand, evidenced by physical presence of food. Availability of food however is no guarantee of food security. People need to own this food or have financial means to purchase the food. This entitlement is what accessibility dimension is all about. Food can be available and people are entitled to food but: what is the guarantee that this will always be so? If people at some period are susceptible to food shortages, then food stability is lacking. Utilization is all about food providing essential food nutrients. Now food insecurity is when at least one dimension from the four is compromised. What is the starting point in tracking food insecurity? Whenever there is an outbreak of food insecurity, food organizations have always responded to availability and accessibility. This research shall takes the same route. Now that the term food insecurity has been contextualized, review of related works can follow.

In 2012, hurricanes hit the United States. Security Homeland [28], describes how postings in social networks in regard to this disaster aided in situation management. Red Cross used social media platforms to communicate with affected communities by responding to the questions raised and disseminating advice on how to handle certain situations. It also sent messages of encouragement to these emotionally troubled citizens. In earth quake outbreaks of Japan 2011, social networks played a related role as described by authors such as Wilson [38] and Fraustino, et al. [11]. For Japan the situation was worse; the internal communication system jammed due to phone call overload. Phones could not go through but citizens could use social networks to communicate with their family and loved members. The story is no different with 2007 California wildfires[31] . For studies such as these describe a limited use of data mining techniques; social networks in this case are mainly acting as a dissemination tool. As a result, there is no way of quantifying the gravity of the problem on various aspects. This is a possible reason why Red Cross reported more than two million posts sent during the hurricane disaster without further detailed analysis [11]. Data mining techniques can reveal richer information as demonstrated in cases of answering questions such as: (1) which pressing issue was the concern of the majority? (2) Was the concern of the majority linked to geo-location of these troubled people? (3) Was there an association between issues raised? Answers to questions like these would attract highly customized actions. Consequently more enhanced services can be provided.

Currently automated techniques are being used to extract meaningful patterns to generate trends. The trends can be compared with baselines if any exists. If none exists, monitoring trends can help in establishing normal trends to provide a basis for comparison. In this way it can be established if the situation is getting beyond normal or still under control. Quantification of information aids in estimation of the required intervention action and this study has attempted it. Studies have emerged to demonstrate how mining social network data can aid in disaster situations. Closest to our study, UN Global Pulse [25] investigated how tweets on food prices relate to official food prices. Using Crimson Hexagon's ForSight software, a classifier was trained to detect each tweet as negative, positive, neutral or confused in relation to food price increase. Incidences of

tweets reporting food price increase were generated into time series. These time series were correlated with actual food price increase from Indonesian official sources. Findings revealed a correlation of 0.4 between food prices from these two sources. No mention of the classifier performance is made. Our study investigates the details of selecting a suitable classifier while reporting its performance which provides more insight. Five classifiers are explored for this purpose and these include; Decision Trees, Support Vector Machines, K-Nearest Neighbors and Neural Networks. UN Global Pulse [25] employed time series for visualization which our study also adopts.

The study also relates to our previous research in [21] which proposed a strategy to harness both phone messages and telephone conversations to predict food insecurity. This is crucial in this era of mobile technology penetration especially in developing world where infrastructure development has been constrained. Mobile technology are playing an instrumental role in promoting sectors such as health, agriculture, and education [12]. In our proposed strategy in [21], focus was on utilizing phone messages and phone conversation– archived for different purposed— for re-use in tracking food insecurity. The study in [21] required users phone messages and telephone conversations which penetrates a lot into users privacy and this complicated access to the necessary data. Taking an option of using tweets (publically freely available) offers flexibility to access the necessary data sets for the current study. In this study we have managed explore performance of data mining tools and some other techniques which were proposed in our study [21] but never experimented.

In the study by Kim and Kim [17], tweets were used to extract people's opinion on nuclear power for Korea. Korea switched to nuclear power as a more affordable alternative. Nuclear power is nevertheless hazardous if not handled carefully. Because of this, Korea had to continuously monitor citizens'opinions on other nuclear plants. This was for consideration prior to Nuclear power installation. The traditional methods originally used to gather people's opinion using survey are labor intensive and expensive. This time Korea chose to use social networks as source of data to track people's opinions. The study considered relevant tweets generated from 2009 to 2013. Like in the study by UN Global Pulse [25], this study exploited time series to visualize trends for the selected period and this provided input for final decisions.

Research by Choi and Bae [7] used an automated system to track and monitor several disasters. The application named Big Data Board was developed. The application examines tweets flowing in per unit time which can be tuned to suit the current situation. It can be tuned to examine tweets in per hour or week or month. The tweets are examined based on selected key words indicative of disasters. The system establishes normal trends based on previous flows. Abnormalities detected trigger alerts. This happens in real time and as a result users have opportunities to take action early enough. Geo-location source of the tweet is also tracked and this is useful in determining where to direct intervention if the need arises. The system can be adopted in the context of our current study. According to Skelly and Smythe [30], Uganda is among the countries with low tweet volume generated by its citizens and this low tweet volume may not be adequate to generate trends such as these. Therefore before employing such a sophisticated system, we chose to explore the described alternative of soliciting more tweets.

Zin, et al. [41] had observed that disaster incidences in social networks are reported in data formats such as text conversations, pictures and videos yet these trends are visualized separately according to the data format. An innovation to visualize trends combining several data formats for the same disaster is proposed and this is expected to provide an enriched visualization. This innovation has been explored using the developed framework employing Markov chaining using two modules; Information Analyzer Module and Event Analyzer Module [41]. Our study deals with only text data, therefore we did not explore the described innovation which required manipulating several data formats.

3 METHODOLOGY

3.1 Data-mining process

A text mining process was formulated to provide a step-by-step process to extract patterns from tweets. A simplified text mining process is a three process step; (1) acquiring text data, (2) mining (extracting) patterns, and (3) generating patterns. Figure 1 shows a simplified text mining process. These steps are often sub-divided into sub-steps for further refinement. Figure 2 shows a representation of a more refined text mining process which was adopted in this study. Preprocessing (a text reformatting activity) is added to smoothen text mining. This includes several sub-tasks: data cleaning, tokenization, stop-word-removal and stemming. Pattern extraction has been expounded to include labeling, classifier training, classifier evaluation, and classifier selection.

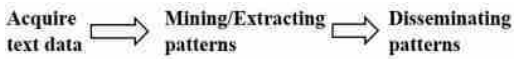

Figure 1: Shows a simplified text mining process

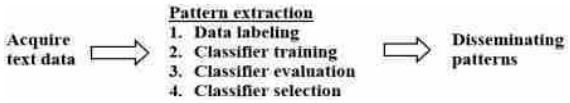

Figure 2: Shows a refined text mining steps

3.2 Acquisition of text

The study developed a script to retrieve the required tweets from Twitter advanced web environment https://twitter.com/search-advanced?lang=en. Tweets were searched from this url using three sets of keywords: (1) food Uganda; (2) hunger Uganda; and (3) famine Uganda. Tweets with words food, hunger and famine can potentially give conversations on food insecurity. We could have used these words in combination but we were interested on how classifiers would behave with datasets for each keyword considered separate. We used these words in pair with the keyword Uganda so that we could retrieve tweets focusing on Uganda. Other set of keywords for example drought Uganda would also provide tweets with conversations on food insecurity. To make

Table 1: Summarizing tweets retrieved

Keywords used	2016	2015	2014
food uganda	2238	1875	1668
hunger uganda	189	79	73
famine uganda	444	308	331

this work manageable under available logistics and time constraints, the study opted to explore with these selected keywords. For each tweet, the script extracted three attributes: the sender, the tweet title and the message. The tweets were for years 2014, 2015, and 2016 as summarized in the Table 1.

3.3 Preprocessing

This stage includes; cleaning tweets, tokenization, stop-word-removal and stemming. The need to clean the tweets was evident; tweets retrieved were mixed with unnecessary text as illustrated in Figure 3. A script for tweet cleaning was developed. The retrieved tweets were saved in comma separated value (csv) format for easy data manipulation. Tokenization was applied to splits text into a stream of words [15]. A token is a basic unit for pattern extractions. Schemes on tokenization take form of monogram (1-gram), bigram (2-gram), trigram (3-gram) to several grams (n-gram). In this study experimentation was carried out to identify which n-gram was optimal using a developed script. Stop-word-removal aims at eliminating irrelevant words computationally containing no essential signal. Prepositions serve as stop-words examples, they do not semantically contribute to the text meaning. Stemming reduces words related to roots to their stems. For example claps, clapped, clapping can be reduced to their stem clap. Stemming reduces vocabulary redundancy and this reduces; dimensionality, computation overhead and noise influence.

DATE	SENDER	TWEET
24 Dec 2014	XXXXX	RECIPE: How to make #Chapati - the most famous fast #food in #Uganda http://wp.me/puKmE-7v\xa0 @tmsruge hope this helps.
27 Dec 2014	XXXXX	FOOD FOR THOUGHT @causticbob: So Uganda has made the penalty for homosexuality life in prison...\r\nI fail to see to the punishment aspect

Figure 3: showing a sample of two unclean tweets. User identities were removed for privacy concerns

3.4 Pattern extraction

In the context of this study, pattern extraction entailed: text labeling, classifier training, classifier evaluation, and classifier selection. In labeling, each tweet retrieved was manually annotated with code 1 if its conversation was about food scarcity and code 0 if not. The text was now trained with classifiers. Previous investigations have identified ten top Machine Learning algorithms [40],[29]. Due to resource constraints, the study explored five classification algorithms (also called classifiers): (1) K-Nearest Neighbors; (2) Support Vector Machine; (3) Decision Trees; (4) Naive Bayes; and (5) Neural Networks.

KNN is a lazy classifier because it memorizes the training sets. It determines a class of unknown instance by searching for a closest instance in the training sets. Now the class of the closest instance is assigned to the unknown instance. Several schemes are used to define closeness among instances. Euclidean distance is the commonest metric used because it is the simplest to compute. To reduce outlier effect, the KNN considers the dominant class among n-nearest neighbors closest to instances of unknown class. KNN suffer from a challenge of computational overhead due to extra space required to hold training example even after fitting the model. Secondly KNN has tendencies of high demand for computation power as distance metric computed involves all the training set. Decision Trees employ decision logic easy for human understanding and as such they are described as white box models. If precautions not taken, Decision Trees are vulnerable to over fitting [6], [20]. According to Kotsiantis [19] early stopping in training and pruning can address over fitting challenge in Decision Trees. Naive Bayes classifier is probabilistic because it computes class probabilities to which a record belongs. It assigns a record to a class with highest probability. While Naive Bayes is based on unrealistic assumption of independence among features, experiments have shown its dependable performance. Support Vector Machine (SVM) is one of the robust and accurate algorithms because it discriminates classes using hyper-plane that maximizes the separation margin between the classes. SVM suffers from situations such as non-linear relation between features and labels. Strategies have to be applied to handle situations such as these. Artificial Neural Networks mimic biological neural networks. Sophisticated Neural Networks can have several layers and this is the reason why they are good for non-linear classification problems. Neural networks suffer from a major drawback of high computational power demand.

The five classifiers were trained to identify whether a tweet is about food insecurity or not. Using evaluation metrics of Receiver Operator Curve Area Under the Curve (ROC-AUC), the best performing classifier was selected for further experimentation. All tweets about food insecurity for each were bundled together ready for generation of trends.

K-fold validation was employed during model training and performance evaluation. This requires dividing the data set into K subsets. K-1 subsets are for training and the remaining single subset is for testing. The criterion for selecting k is not well defined. Any value K=5 and above is considered fair. Studies are fond of using K=5 and K=10. Some heuristics recommend big values of k for huge datasets. The data set was partitioned into K=5 equal-sized subsets. Four subsets were for training and one subset was for testing. The process was repeated five times but for each manipulation a different single subset was for testing. Area Under the Curve of Receiver Operator Characteristic (ROC-AUC) for each subset testing was computed. ROC-AUC was a suitable choice as some data sets available for the study had unbalance distribution of training examples between classes. Overall ROC-AUC was the average for accuracies computed from all test subsets.

3.5 Generating food insecurity trends

Datasets for 2014 and 2015 were subjected to supervise learning to identify the most performing classifiers. The best performing classifiers were employed to automatically label 2016 twitter data sets. Monthly counts for tweets with conversations on food each for year were computed. Trends in form of time series graphs for both classifier and human labeled twitters were generated for visualization and interpretation. Correlation between these two trends was assessed. Twitter trends were also compared with trends using official alternative sources. Now correlation between these two trends was computed to establish validity of trends generated from tweets.

All twitter datasets for 2015 were combined aggregated. The same was done for datasets of 2016. Food price inflations were used as alternative official sources. These were from Uganda National Bureau of Statistics [34], [35]. Monthly tweets counts reporting food insecurity and food price inflation values were normalized to create uniform scaling for easy graphic plotting and assessment. Three normalization techniques were reviewed and these include Z-score normalization, Decimal scaling normalization, and Min-Max normalization [27]. Z-score aligns data point towards the mean. It measures how data scores deviate from the mean value. This was not appropriate for the study. Decimal scaling normalization maps data values by moving decimal point of values of feature say X. The study used Min-Max normalization which is a linear and a simple technique to implement. Data was fit into pre-defined boundary and [0, 1] was chosen as the boundary.

4 RESULTS

4.1 Preprocessing

Table 2 shows options that yielded optimal performance during pre-processing stage.

Table 2: Showing preprocessing optimal choices

Preprocessing stage	Optimal parameter
Stop-Word removal	Document frequency of 0.6 and above
n-gram choice	Combination of 1-gram and 2-gram

4.2 Classifier performance

Two data sets produced reliable performance (see Table 4 and Table 6). These classifiers were above the random classifier. The other two tables, best classifier performances were worse than random classifier. In Table 3, the best classifier performance produced 0.819 compared with random performance of 0.892. This is a deviation of 0.073 (7 percent) below expected minimum performance. In Table 5 the best classifier performance was 0.901 compared to 0.934. This is a deviation of 0.033 (3 percent) below expected minimum performance. Now clearly worst performance was in Table 3. This was for tweets retrieved with keywords food uganda.

Table 3: Showing performance for tweets retrieved with keywords food Uganda operating on random classifier of 0.892

CLASSIFIER	ROC-AUC score
KNN	0.628
N-BAYES	0.675
SVM	0.819
D-TREES	0.709
N-NETWOKS	0.762

Table 4: Showing classifier performance for tweets retrieved with keywords hunger Uganda operating on random classifier of 0.503

CLASSIFIER	ROC-AUC score
KNN	0.576
N-BAYES	0.775
SVM	0.720
D-TREES	0.624
N-NETWOKS	0.745

Table 5: Table showing classifier performance for tweets retrieved with keywords famine Uganda operating on random classifier of 0.934

CLASSIFIER	ROC-AUC score
KNN	0.500
N-BAYES	0.901
SVM	0.746
D-TREES	0.496
N-NETWOKS	0.710

Table 6: Table showing classifier performance for all combined tweets operating on random classifier of 0.798

CLASSIFIER	ROC-AUC score
KNN	0.662
N-BAYES	0.807
SVM	0.853
D-TREES	0.748
N-NETWOKS	0.837

4.3 Classifier trends against manual trends

Food insecurity trends for classifier labeled tweets are compared with the trends for human labeled tweets. This was done on the two datasets where classifiers had generated reliable performance. The correlation in the combined datasets (shown in Figure 4) was better in the dataset retrieved with keywords hunger + Uganda (shown in Figure 4).

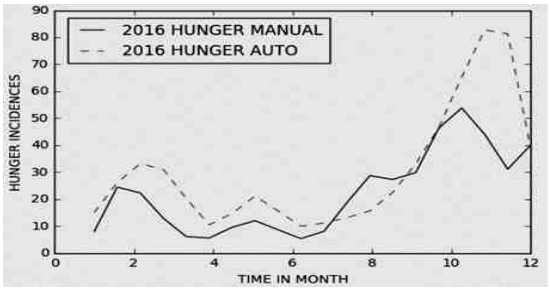

Figure 4: Showing trends for hunger tweets for human (manual) label against classifier (automated) generated with computed correlation of 0.804

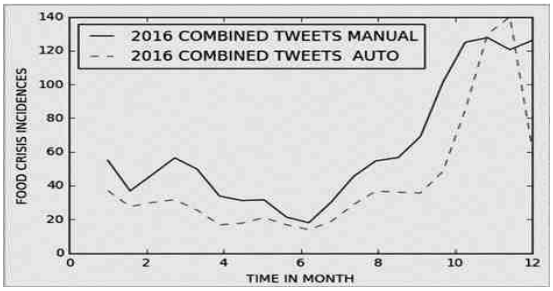

Figure 5: Showing trends for combined tweets for human (manual) label against classifier (automated) generated with computed correlation of 0.848

4.4 Comparing food insecurity with other sources

Twitter trends against UBOS (extracted from [34], [35]) trends both on food crisis. For the two years (2015 and 2016) the correlation for the year 2016 is higher than for 2015 as shown in the Figures 6 and 7.

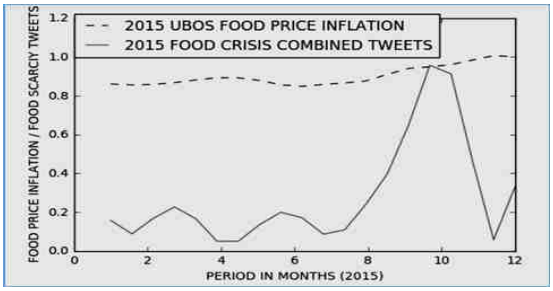

Figure 6: Showing food insecurity trends generated from tweets compared with trends from food price inflation for 2015. Computed correlation between trends is 0.35

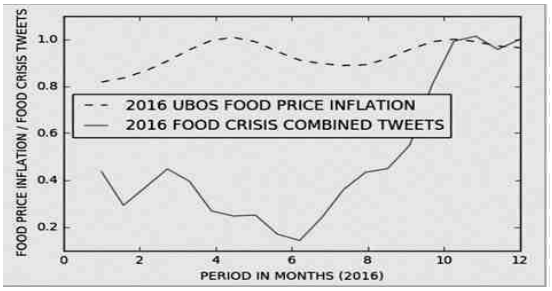

Figure 7: The figure below showing food insecurity trends generated from tweets compared with trends from food price inflation for 2016. Computed correlation between trends is 0.57

5 DISCUSSION

We found that we can generate trends of food insecurity in Uganda using tweets. Our work relates with a study by UN Global Pulse [25] which demonstrated use of tweets reporting food prices to predict actual food price increase. Our study however demonstrates use of tweet conversations to track food insecurity. Our study therefore provides an alternative option to study food challenge. In our study tweets were from all over the world while in UN Global Pulse all tweets were coming from citizens. We opted to use tweets from all over the world due to low volume tweet that could not aid in the study. The study has demonstrated an alternative option to solicit tweets for our nation with low volume of tweets from citizens. By generating these trends, the research demonstrated how it addressed research question one of tracking food insecurity from tweets. This is possible as these trends give a general picture how food insecurity varies over time. This section has also partially addressed research question two by showing how this work relates to research done by UN Global Pulse [25]. The strategy used in this work can be tried in regions such as Kenya facing similar situation (low tweet volume from the citizens). What remains unanswered: How reliable could such tweets be compared with tweets from the local citizen? This is an interesting investigation perhaps in a separate study.

The best classifier performances are comparable with human judgment in classifying some tweet (see Table 4 and Table 6).This demonstrates the extent to which research question three was addressed on how to automate classification tweets. This means human intervention of sorting relevant tweets can be eliminated for these datasets (shown in Table 4 and Table 6). We now again demonstrate how research question two was addressed. Automation of text classification demonstrated in this work relates to studies that employed data mining to; (1) predict whether stock prices will increase or not using news articles [18], and (2) predict sentiments of user comments [5]. The difference with our study is what is being predicted but the concept is the same; applying data mining algorithms to make prediction. The uniqueness brought forth in this study is that to the best our knowledge we have not come across a study applying data mining classifiers as this study has demonstrated. Other general applications of data mining algorithms

include Information Retrieval, Document Summarization, and Recommender Systems. So to say data mining opens a wide range of opportunities to explore.

In Table 3 and Table 5 overall classifier performances are low; worse than random classifiers. This reveals the extent to which research question three could not be addressed of automatically classifying the tweets as planned. What is the cause? Language is complex characterized by many ambiguities such as sarcasm, idioms, and contextual meaning. So classifiers work against this complexity. This partially accounts for observed low performance. However inherent language complexities have not prohibited success of classifiers in investigations such as those described in [16], [26] and [5]. Therefore language complexities is not sufficient to explain low performances observed in these tables. Let us focus first on the Table 3. We have assumed that tweet conversations on food strictly belong to two classes; either a tweet is talking about food insecurity or not. Tweets with food as key word have high potential to communicate more than two levels of food security. For example five classes are possible: (1) high food crisis, (2) moderate food crisis, (3) neutral situation, (4) moderate food abundance, and (5) high food abundance. Tweets on food were forced into two categories against other possible categorical diversities. Now the classifiers are being forced to squeeze tweets between two classes; belonging to either food insecurity class or not. This simplification is perhaps is dis-stabilizing classifier learning capacity. On contrary presence of a word hunger or famine contains strong signal to communicate food insecurity or not. Therefore for tweets either retrieved by hunger or famine would portray a better performance. This is the cause for promising performance (shown in Table 4) with tweets retrieved with hunger however for tweets retrieved with famine this was not the case. What is the cause of unexpected poor performance shown in Table 5 ? Examining tweets retrieved with famine, it was found that the word famine was sometimes sarcastically used yet these classifiers do not always deal well with this situation. Secondly, tweets containing famine as a keyword had unbalanced datasets distributed between the two classes and according to Ali et al. [2] class imbalance is a big threat to classifier learning. Several techniques (for example see [2]) have been proposed to circumvent this. Exploring these techniques is worth trying perhaps in a future venture.

Let us now examine optimal options observed at pre-processing stage (as shown in Table 2). This was for optimizing classifier performances to indirectly support achievement of research question three. Stop word removal using a standard dictionary sklean library was unhelpful. For some domains standard dictionary of stop-words is not appropriate and applying customized/domain dictionary of stop words is a better option. Customized dictionary was appropriate for this study. Removing words with 60 percent and above document frequency in the tweets was optimal. Empirical evidence shows threshold varies from situation to situation. Removal of low frequency word was unhelpful. Sophisticated schemes to generate stop-words have been proposed. These include use of; (1) backward filter-level performance [22] and (2) Poisson distribution [4]. Exploring performance of schemes such as these can be an interesting pursuit. On ngram choice, 1-3 ngram values were varied. A combination of a monogram (1-gram) and bigram (2-gram) was optimal. Single gram are basic unit for signal/meaning but in language meaning can sometimes be delivered best with a combination of grams. Our study is such scenario that requires more than a monogram.

Tweet conversations on food insecurity are highest from September to December (shown in Figure 4 and Figure 5) and this compelled the study for investigation. This food insecurity intensity agrees with media reports of two Ugandan news media; Monitor (www.monitor.co.ug) and New Vision (www.newvision.co.ug). For these media we searched for articles reporting major incidences of food insecurity for years 2015 and 2016. We found that highest cases relating to food insecurity in Uganda also included this periods. This generated credibility for the trends from these tweets.

The generated food insecurity trends were comparable with trends from food price inflations of UBOS (see Figure 6 and Figure 7). This was to address research question four of ensuring validity of trends generated. For years 2015 and 2016 correlations between these trends were 0.34 and 0.57 respectively. Statistically 0.34 is low and 0.57 is average correlation. This is a promising correlation as far validity is concerned. Better correlations are a possibility with sophisticated algorithms such as ensemble algorithms. Secondly failure to obtain high correlation can be due to differences on how food inflation directly map to food insecurity incidences. Use of other data sets can improve level of correlation and this requires investigation. These results are comparable with related work of Pulse UN Global [25]. The correlation of 0.4 obtained by UN Global Pulse [25] relating tweet food prices with official sources is comparable with correlations (of 0.35 and 0.57) obtained in this study.

6 CONCLUSION AND FUTURE WORK

The study has generated empirical evidence showing use of tweets to provide proxy information about food insecurity in Uganda. This is in the context of low tweets generated by citizens. The study has further illustrated use of classification algorithms to automatically filter relevant tweets for generating food insecurity trends. The trends from this study can reveal an emerging incidence of food insecurity. Food security agencies like Famine Early Warning System in Uganda can use this information to plan for appropriate action. On a wider perspective, this research belongs to investigations that are mining useful patterns from web data (social network conversations inclusive) to enhance decision making and policy formation. For performance improvement on this work, future work can consider: (1) aggregating tweets with other datasets, (2) applying ensemble algorithms, and (3) using unexplored algorithms. The study can also be extended to countries experiencing similar context of low tweet volume generated from the citizens. Lastly, exploring other situations other than food insecurity using the strategy described is a possibility.

REFERENCES

[1] Mohammad-ali Abbasi, Shamanth Kumar, Jose Augusto, and Andrade Filho. 2012. Lessons Learned in Using Social Media for Disaster Relief - ASU Crisis Response Game. In *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer-Verlag, Berlin, 282–289. http://www.public.asu.edu/{~}huanliu/papers/SBP12Game.pdf[Retrievedon30thMay2016]

[2] Aida Ali, Siti Mariyam Shamsuddin, and Anca L Ralescu. 2015. Classification with class imbalance problem : A Review. *Int. J. Advance Soft Compu. Appl* 7, 3 (2015), 176–2014.

[3] Godfrey B a Bahiigwa. 1999. *Household Food Security in Uganda: an Empirical Analysis.* Technical Report 1-24.

[4] Vicenç Parisi Baradad and Alexis-michel Mugabushaka. 2015. Corpus Specific Stop Words to Improve the Textual Analysis in Scientometrics. In *15th International Conference of the International Society for Scientometrics and Informetrics (ISSI).* 999–1005. https://ai2-s2-pdfs.s3.amazonaws.com/618d/a4f6d5cd329d3bc498d9457f575cbdfaf53d.pdf[Accessedon3rdDec2017]

[5] Aditya Bhardwaj, Yogendra Narayan, and Maitreyee Dutta. 2015. Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty. In *4thInternational Conference on Eco-friendly Computing and Communication Systems*, Vol. 70. Elsevier Masson SAS, 85–91. https://doi.org/10.1016/j.procs.2015.10.043

[6] Rich Caruana and Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*. 161–168. https://doi.org/10.1145/1143844.1143865

[7] Seonhwa Choi and Byunggul Bae. 2015. The real-time monitoring system of social big data for disaster management. *Lecture Notes in Electrical Engineering* 330 (2015), 809–815. https://doi.org/10.1007/978-3-662-45402-2_115

[8] Deloitte. 2015. *Facebook ' s Global Economic Impact.* Technical Report. http://www2.deloitte.com/content/dam/Deloitte/uk/Documents/technology-media-telecommunications/deloitte-uk-global-economic-impact-of-facebook.pdf[Accessedon30thMay2016]

[9] FAO, IFAD, and WFP. 2014. *The State of Food Insecurity in the World.* Technical Report.

[10] FEWS NET Uganda. 2009. *Uganda Food Security Outlook.* Technical Report July to December 2009.

[11] Julia Daisy Fraustino, Liu Brook, and Jin Yan. 2012. *Social Media Use during Disasters.* Technical Report. https://www.start.umd.edu/sites/default/files/files/publications/START{_}SocialMediaUseduringDisasters{_}LitReview.pdf[Accessedon12thJune2016]

[12] Ahmed Imran, Val Quimno, and Mehdi Hussain. 2016. Current landscape and potential of mobile computing research in the least developed countries. *The Electronic Journal of Information Systems in Developing Countries* 74, 5 (2016), 1–25. https://doi.org/10.1002/j.1681-4835.2016.tb00539.x

[13] IPC. 2017. *Uganda - Current Acute Food Insecurity Situation.* Technical Report March.

[14] JLINK LABS. 2016. *Data Provenance as a Service.* Technical Report. http://www.jlinclabs.com/wp-content/uploads/2016/09/JLINC{_}WP{_}DataProvenanceAsAService{_}20160912.pdf[Accessedon30thMarch2017]

[15] Ayman E. Khedr, S. E. Salama, and Nagwa Yaseen. 2017. Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. *International Journal of Intelligent Systems and Applications* 9, 7 (2017), 22–30. https://doi.org/10.5815/ijisa.2017.07.03

[16] Dong Sung Kim and Jong Woo Kim. 2014. Public Opinion Sensing and Trend Analysis on Social Media : A Study on Nuclear Power on Twitter 1. *International Journal of Multimedia and Ubiquitous Engineering* 9, 11 (2014), 373–384. http://onlinepresent.org/proceedings/vol51{_}2014/51.pdf[Accessedon12thJune2016]

[17] Dong Sung Kim and Jong Woo Kim. 2014. Public Opinion Sensing and Trend Analysis on Social Media: A Study on Nuclear Power on Twitter. *International Journal of Multimedia and Ubiquitous Engineering* 9, 11 (2014), 373–384. http://www.sersc.org/journals/IJMUE/vol9{_}no11{_}2014/36.pdf[Accessedon27thMarch2017]

[18] Yoosin Kim, Seung Ryul Jeong, and Imran Ghani. 2014. Text Opinion Mining to Analyze News for Stock Market Prediction. *International Journal of Advances in Soft Computing and Its Applications* 6, 1 (2014), 1–13. http://home.ijasca.com/data/documents/Paper-ID-424-IJASCA{_}Formated.pdf[Acessedon12thDec2016]

[19] Sotiris B. Kotsiantis. 2007. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* 31, 2007 (2007), 249–268. https://doi.org/10.1115/1.1559160 arXiv:http://www.informatica.si/index.php/informatica/article/view/148

[20] Praful Koturwar, , Sheetal Girase, and Mukhophadhyay Debajyoti. 2014. A Survey of Classification Techniques in the Area of Big Data. *International Journal of Advance Foundation and Research in Compute* 1, 11 (2014), 1–7. arXiv:1503.07477 https://arxiv.org/ftp/arxiv/papers/1503/1503.07477.pdf[Accessedon16thJan2018]

[21] Andrew Lukyamuzi, John Ngubiri, and Washington Okori. 2015. Towards Harnessing Phone Messages and Telephone Conversations for Prediction of Food Crisis. *International Journal of System Dynamics Applications* 4, 4 (2015), 1–16. https://doi.org/10.4018/IJSDA.2015100101

[22] Masoud Makrehchi and Mohamed S. Kamel. 2017. Extracting domain-specific stopwords for text classifiers. *Intelligent Data Analysis* 21, 1 (2017), 39–62. https://doi.org/10.3233/IDA-150390

[23] Lawrence Phillips, Chase Dowling, Kyle Shaffer, Nathan Hodas, and Svitlana Volkova. 2017. *Using Social Media to Predict the Future: A Systematic Literature Review.* Technical Report. 1–55 pages. arXiv:1706.06134 http://arxiv.org/abs/1706.06134[Accessedon19thOct2017]

[24] G M Poppy, S Chiotha, F Eigenbrod, C a Harvey, M Honzák, M D Hudson, A Jarvis, N J Madise, K Schreckenberg, C M Shackleton, F Villa, and T P Dawson. 2014. Food security in a perfect storm: using the ecosystem services framework to increase understanding. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 369 (2014). http://rstb.royalsocietypublishing.org/content/369/1639/20120288.short

[25] Pulse U N Global. 2014. *Mining Indonesian Tweets to Understand Food Price Crises.* Technical Report February. 1–18 pages. http://www.unglobalpulse.org/sites/default/files/Global-Pulse-Mining-Indonesian-Tweets-Food-Price-Crisescopy.pdf[Accessedon26thMarch2017]

[26] Mark Sabini, Gili Rusak, and Brad Ross. 2017. *Understanding Satellite-Imagery-Based Crop Yield Predictions.* Technical Report. Stanford University. http://cs231n.stanford.edu/reports/2017/pdfs/555.pdf[Accessedon23thOct2017]

[27] C. Saranya and G. Manikandan. 2013. A study on normalization techniques for privacy preserving data mining. *International Journal of Engineering and Technology* 5, 3 (2013), 2701–2704. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.411.1996{&}rep=rep1{&}type=pdf[Accessedon2ndDec2017]

[28] Security Homeland. 2014. *Using Social Media for Enhanced Situation Awareness and Decision Support.* Technical Report. 1–44 pages. https://www.hsdl.org/?view{&}did=755891[Accessedon12thJune2016]

[29] Nesma Settouti, Mohammed El Amine Bechar, and Mohammed Amine Chikh. 2016. Statistical Comparisons of the Top 10 Algorithms in Data Mining for Classi cation Task. *International Journal of Interactive Multimedia and Artificial Intelligence* 4, 1 (2016), 46–51. https://doi.org/10.9781/ijimai.2016.419

[30] Katie-rose Skelly and Sabelle Smythe. 2016. Learning Food Price Indicators from Twitter Data. (2016).

[31] Jeannette Sutton, Leysia Palen, and Irina Shklovski. 2008. Backchannels on the Front Lines : Emergent Uses of Social Media in the 2007 Southern California Wildfires. In *Proceedings of the 5th International ISCRAM Conference âĂŞ Washington, DC, USA, May 2008.* Washington, DC, USA. http://www.jeannettesutton.com/uploads/BackchannelsISCRAM08.pdf[Accessedon12thJune2016]

[32] The World Bank. 2016. *Poverty Assessment Report 2016.* Technical Report.

[33] Uganda Bureau of Statistics. 2007. *Projections of demographic trends in Uganda 2007-2017. Volume I.* Technical Report December.

[34] Uganda Bureau of Statistics. 2015. *Low season report CONSUMER PRICE INDEX DECEMBER 2014.* Technical Report. Uganda Bureau of Statistics, Kampala. http://www.ubos.org/onlinefiles/uploads/ubos/cpi/cpiDec2015/FINALCPIReleaseDecember2015.pdf[Accessedon6thNov2017]

[35] Uganda Bureau of Statistics. 2017. *Uganda consumer price index: 2009/10=100.* Technical Report. Uganda Bureau of Statistics, Kampala. 1–37 pages. http://www.ubos.org/onlinefiles/uploads/ubos/cpi/cpidec2016/CPIPublicationforDecember2016.pdf[Accessedon7thNov2017]

[36] Marcela Umana-Aponte. 2011. *Long-term effects of a nutritional shock: the 1980 famine of Karamoja, Uganda.* Technical Report. 60 pages pages.

[37] Charith Wickramaarachchi, Alok Kumbhare, Marc Frincu, Charalampos Chelmis, and Viktor K Prasanna. 2015. Real-time Analytics for Fast Evolving Social Graphs. In *Cluster, Cloud and Grid Computing (CCGrid), 2015 15th IEEE/ACM International Symposium on.* IEEE, Shenzhen, 829 – 834. https://ganges.usc.edu/svn/pg/pubs/preprint/Charith-Scale-2015.pdf

[38] Jennifer Wilson. 2012. *Responding to Natural Disasters with Social Media : A Case Study of the 2011 Earthquake and Tsunami in Japan by.* Ph.D. Dissertation. Simon Fraser University.

[39] World Food Summit. 1996. *The Rome declaration on world food security.* Technical Report. 14–17 pages. https://doi.org/10.2307/2137827

[40] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J Mclachlan, Angus Ng, Bing Liu, Philip S Yu, Zhi-hua Zhou, Michael Steinbach, J Hand David, and Dan Steinberg. 2007. Top 10 algorithms in data mining. *Knowledge and Information Systems* 4, 1 (2007), 1–37. https://doi.org/10.1007/s10115-007-0114-2

[41] Thi Thi Zin, Pyke Tin, Hiromitsu Hama, and Takashi Toriu. 2013. Knowledge based Social Network Applications to Disaster Event Analysis. In *Proceedings of the International MultiConference of Engineers and Computer Scientists 2013*, Vol. I. Hong Kongo, China. http://www.iaeng.org/publication/IMECS2013/IMECS2013{_}pp279-284.pdf