An Ngram-Based Approach to Determine Trends and Patterns in the Social Networks

Constance Mukina Ngila
Department of Information Technology,
Jomo Kenyatta University of
Agriculture and Technology
Jomo Kenyatta University of
Agriculture and Technology
Nairobi, Kenya
cmngila@gmail.com

Waweru Mwangi
Department of Information Technology,
Jomo Kenyatta University of
Agriculture and Technology
Jomo Kenyatta University of
Agriculture and Technology
Nairobi, Kenya
waweru_mwangi@icsit.jkuat.ac.ke

Michael Kimwele
Department of Computing,
Jomo Kenyatta University of
Agriculture and Technology
Jomo Kenyatta University of
Agriculture and Technology
Nairobi, Kenya
mkimwele@jkuat.ac.ke

***Abstract*— The recent progress in computing has made iteasier to collect and store huge amounts of information in a text. The growing size of text datasets in text mining and the high dimensionality associated with knowledge discovery is a great challenge that makes it difficult to classify documents into various categories and sub- categories. This paper focuses on how text can be mined from social networks and then categorized using n-grams to determine specific trends and patterns. The main aim of Knowledge Discovery is to extract knowledge from data in the context of large databases. The volume of information that is available is increasing every day. This data ranges from that used in business transactions to scientific data, sensor data, pictures, videos, etc. There is, therefore, a need for a system capable of extracting the core of available information and automatically generating reports, opinions, or summaries of data to aid organizations in better decision-making. Knowledge Discovery is a repetitive process where evaluation measures are often enhanced, mining done on data can be refined, there is an integration of new data, and the data is transformed to get accurate and more appropriate results. The data collected from social networks need to be filtered to capture specific text that will be useful to a PR brand following what clients say about their products online. There is a need for a technique that will provide a quick and precise way of fetching specific text from huge amounts of data on social networks to help analyze the feedback. This research analyzes the use of n-grams to fetch specific text from near-real-time customer feedback that is in the form of large data on Twitter to help Public Relations agencies determine the trends and patterns that will help them align their brands with customer preferences.**

Keywords—knowledge discovery, data mining, trends, and patterns.

I. INTRODUCTION

The capability to gather data in various instruments, devices, and formats that are both from independent and connected applications has considerably overtaken our ability to process, evaluate, store, and comprehend these data sets. Social networking has thrived, and platforms like Facebook, Instagram, and Twitter extensively allow users to create content freely, further magnifying the already massive volume of web data. Knowledge Discovery is an area that focuses on methodologies that can be used for extracting meaningful and useful knowledge from data [1]. The ongoing speedy growth of online data from theInternet and the extensive use of databases has created a massive need for knowledge discovery methodologies [2]. The challenge of mining knowledge from data has led to research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing, to convey advanced BI and solutions for web discovery [3]. Social networking was initially intended for informal communication, but the growthof social media has seen people and companies conducting formal meetings online [4]. As opposed to an ecommerce website, which is an entirely formal means ofcommunication, networking through social platforms iscomparable to conversing with a person face to face but through an online application like Facebook or Twitter [5]. People and devices have become moreand more connected. Huge numbers of such connected components have generated a huge data ocean, and meaningful information needs to be discovered from the large data to help improve quality of life [6]. The use of n-grams in knowledge discovery simplifies the complexities of huge datasets by categorizing and classifying data according to sentiments. N-grams help fetch specific text by capturing wordsand characters, identifying relationships in the groups of data, and providing valuable insights into underlying data patterns.

II. JUSTIFICATION FOR THE STUDY

The growing volume of text data presents the need to implement techniques that help classify data, and this paper examines the use of data mining techniques to classify and categorize text data from social networks. Through knowledge discovery, it is possible to achieve insight into trends and patterns that Public Relations agencies can employ to align their brands according to feedback. N-grams help identify trends and patterns to understand customer sentiments and feedback further.

III. PRELIMINARIES

A. N-grams

In the fields of computational linguistics and probability, an n-gram has been defined as a contiguous sequence of n items from a given sample of text or speech

2023 IEEE AFRICON | 979-8-3503-3621-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/AFRICON55910.2023.10293617

979-8-3503-3621-4/23/$31.00 ©2023 IEEE

[7]. They are a sequence of N units or tokensof text where those units are typically single characters orstrings that are delimited by spaces. According to the application, the items can be syllables, letters, words, or base pairs. The n-grams are typically collected from a text or speech corpus [8]. According to Banerjee and Pedersen, n-grams are a simple representation that suits written languages. Therefore, they are a sequence of N units or tokens of text where those units are typically single characters or strings that are delimited by spaces [9].

B. Computing N-grams

When computing the n-grams, one word is moved forward (although several words can be moved forward in other scenarios). For example, in this sentence: "The quick brown fox jumped over the lazy dog." If N=2 (bigrams), then the n-grams would be:

- the quick
- quick brown
- brown fox
- fox jumped
- jumped over
- over the
- the lazy
- lazy dog

There are eight n-grams in this case. It moves from the ->quick to quick->brown to brown->fox, etc. One word is usually moved forward in order to generate the next bigram.

IV. METHODOLOGY

A. Twitter API

Data for this study was collected on Twitter. Tweets containing specific hashtags and posted within a particular timeframe were extracted. Users on Twitter post unfiltered opinions that can be retrieved with ease. The method applied here was web scraping of the Twitter API using the Python Library called Tweepy, which enabled access to the API. Web scraping is the method that was used to collect large amounts of information from the website.

Web scraping has been described as an automated method used to extract huge amounts of data available on websites. The data that is available on the websites is typically unstructured. Therefore, web scraping was implemented in this study to collect unstructured data and further store it in a structured form.

B. Data Analysis

Using the search parameters "id" and "count," Tweepy provides access to users' handles and the maximum number of recent tweets intended to be scraped from the timeline. Below is an example of how a twitter user's timeline was scraped to retrieve their most recent 50 tweets.

```
\begin{algorithm}[H]
\caption{Twitter Data Retrieval}
\label{alg:twitter_retrieval}
\begin{algorithmic}[1]
\State $\text{username} \gets \text{'jack'}$
\State $\text{count} \gets 150$
\State \textbf{try}:
\State \quad $\text{tweets} \gets
\text{tweepy.Cursor(api.user_timeline,
id=username).items(count)}$
\State \quad $\text{tweets_list} \gets
[[\text{tweet.created_at}, \text{tweet.id}, \text{tweet.text}]
\text{ for tweet in tweets}]$
\State \quad $\text{tweets_df} \gets
\text{pd.DataFrame(tweets_list)}$
\State \textbf{except} $\text{BaseException as e}$:
\State \quad $\text{print('failed on_status,', str(e))}$
\State \quad $\text{time.sleep(3)}$
\end{algorithmic}
\end{algorithm}
```

The program above begins by creating a query method using the "id" and "count" parameters. A tweets list is then created to pull information from tweets iterable object. The next step is to create a data frame from that particular tweets list, where it is possible to add or remove columns as tweet information is removed. The search can be further optimized by adding more parameters like "exclude_replies", "include_rts","trim_user" among others.

C. Preprocessing

Data preprocessing is done to improve the quality of the data. The data undergoes cleaning, then normalizing, then transforming, and finally extracting all relevant features from raw data [10].

Data preprocessing improves the overall performance and presentation of machine learning algorithms, which results in accurate data mining.
Performing successful knowledge discovery from irrelevant, redundant, and noisy data requires accurate identification of extreme data values, and then filling up all the missing values present many challenges.

Data Cleaning is conducted to fix or remove incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data from the data set. If data is incorrect, then algorithms and outcomes are unreliable. This process of data cleaning recognizes the partial, incorrect, imprecise, or inappropriate parts of data from datasets. Data Cleaning may ignore tuples that contain missing values, or it could alter the values compared to an already-known list of entities. This ensures that the data will become consistent with all the other data sets available in the system.

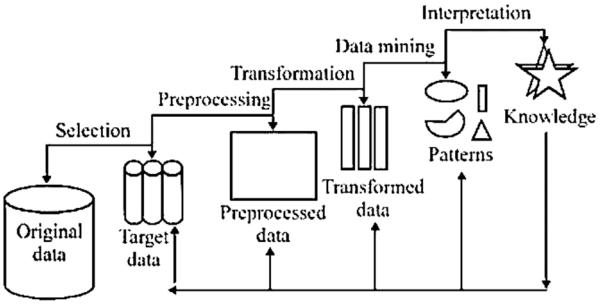

Fig. 1: Data Preprocessing

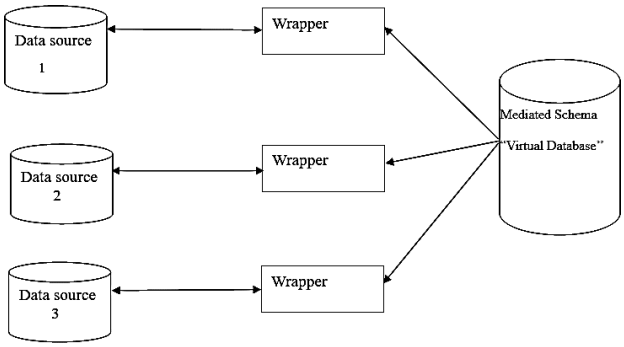

Fig. 2: Data Integration

Data reduction helps obtain a reduced representation of a significantly smaller volume of data while still maintaining the integrity of the original data. This means mining the reduced data volume is simpler and more efficient, producing the same analytical results.

V. EVALUATION AND RESULTS

Through sentiment analysis, the tweets are classified according to respective sentiments by categorizing them as positive, neutral, or negative. Classification of the sentiments helps identify how trends and patterns change over time, and what decisions and predictions can be made based on the sentiments. Importing TextBlob in Python helps conduct sentiment analysis on data sets. The TextBlob library was helpful in this study to gather sentiments from a data set using n-grams. The snippet below helped categorize sentiments about a product as positive, neutral, or positive.

```
 \usepackage[ruled]{algorithm2e}
\begin{verbatim}
\usepackage{python}
\begin{python}
from textblob import TextBlob
from sklearn.feature_extraction.text import
CountVectorizer
from sklearn.naive_bayes import MultinomialNB
reviews = ["This new phone model is really good.",
        "I don't have special feelings about this new
model.",
        "I dislike the update in this new phone model."]
# Labels for sentiment analysis
labels = ['positive', 'neutral', 'negative']
# Create n-grams
ngram_range = (1, 2) # Use unigrams and bigrams
vectorizer =
CountVectorizer(ngram_range=ngram_range)
X = vectorizer.fit_transform(reviews)
classifier = MultinomialNB()
classifier.fit(X, labels)
test_review = "I love the latest features of this phone."
test_X = vectorizer.transform([test_review])
predicted_label = classifier.predict(test_X)[0]
print("Predicted Sentiment: ", predicted_label)
\end{python}
\end{verbatim}
```

VI. SUMMARY

The results of the study revealed that n-grams are an effective technique to determines trends and patterns in the social networks. Applying sentiment techniques with n-grams made it possible to evaluate the polarity of sentiments provided on social media posts, specifically tweets. The findings demonstrated the effectiveness of the n-gram based approach in interpreting trends and patterns contextually. To acknowledge the limitations of the study, it is important to note that the effectiveness of the approach depended on context and the dataset under observation. Future studies could examine the technique when incorporating factors like user demographic, and temporal analysis.

VII. CONCLUSION

This study demonstrated the effectiveness of n-grams in sentiment analysis as a technique to determine trends and patterns in social networks. The results obtained in the study provided valuable insight that organizations like PR agencies can apply when determining feedback that they can implement to align their products with clients' trends.

ACKNOWLEDGMENT

I sincerely thank my supervisors for taking the time to read my work, correcting me, and helping me narrow down my ideas. I wish to thank my parents for their prayers, encouragement, and support.

REFERENCES

[1] L. X. Z. D. Chen Z, "Discovering event trends from social media data: A multi-modal approach," *Knowledge-Based Systems,* pp. 54-64, 2018.

[2] S. Chen, X. Li and C. Wang, "Mining and analyzing event trends in social media data: A survey," *Information Processing & Management,* vol. 58(1), no. 102481, 2021.

[3] M. Johnson and L. Williams, "Predicting trends in social media data using N-gram analysis," *International Journal of Information Management,* pp. 78-92, 2022.

[4] J. Kim and H. Park, "Analyzing event-related tweets for trend detection using N-gram models," *Journal of Information Science,* vol. 47(4), pp. 527-541, 2021.

[5] X. Li and L. Yang, "Analyzing event trends in social media using topic modeling," *Online Information Review,* pp. 1354-1372, 2019.

[6] X. Liu, S. Li and W. Li, " Trend analysis of social media data for event detection and prediction," *Future Generation Computer Systems,* pp. 129-139, 2021.

[7] A. Smith, "Social media analytics for event trend analysis," *Journal of Data Science,* pp. 145-164, 2022.

[8] X. Wang, J. Liu and W. Xu, "Analyzing Twitter data for event trend detection using N-gram models," *IEEE Transactions on Knowledge and Data Engineering,* pp. 249-262, 2019.

[9] Y. Wang, H. Chen and W. Gao, "Event detection and trend analysis in social media data," *Information Sciences,* pp. 160-178, 2020.

[10] Q. Zhang, J. Tang and H. Gao, "N-gram analysis for event trend detection in social media," *Journal of Computational Science,* 2020.