**Analysis on Food Tweets using Natural Language Processing Technique**

 1.Dr. K. Arutchelvan, 2.R. Ramachandran, 3.Dr.G.Vaitheeswaran, 4.Dr.A.Murugan

1,2, Assistant Professors/Programmer, Department of Computer and Information Science,

 Annamalai University, Chidambaram.

3.Organization: Victory Lance, 4.Assistant Professor, Department of Computer Science, Thiru Kolanjiappar Government arts College (Grade-1),Virudhachalam

E-mail: ramachandranr.au@gmail.com,karutchelvan@yahoo.com mailvaithees@gmail.com,muruganapau@gmail.com

**ABSTRACT**

The rapid development of web-based media such as Facebook, Twitter, Quora, and more empowers people to freely share their attitude / opinion / sentiments / insights on the Internet. People look to these sites for other people's evaluations of a product or management. Business associations need them to improve their advertising. This information is shared or disclosed in electronic form such as voice, video, texts etc. Analyzing this data requires an intelligent technique known as Natural Language Processing (NLP). NLP offers a variety of solutions for many text applications. One of the main applications using NLP is concept mining.

Opinion mining (OM) of code-mixed content has varied applications in decision mining, from labeling client surveys to recognizing social or political opinions of subpopulations. In this research work, it presents the Twitter Foodie Corpus (TFC). Contains a collection of tweets in a tight niche identified by food, beverages, eating and drinking. It has around 1 million tweets collected with keywords like biryani, dosa, idli etc. The collected tweets are processed and analyzed using well-known machine learning models such as Naïve Bayes and Multilayer Perceptron (MLP). The nave base outperformed the MLP.

1. **INTRODUCTION**

Quick improvement of web-based media like face book, twitter, fora, and so on empowers people to impart their insights freely on the Web, which can be gotten to by others for their dynamic. People are seeing these locales for assessments of others about an item or administration prior to making a buy or others feelings about political up-and-comers. Business associations need them for improving their promoting. Nonetheless, the serious issue here is that despite the fact that volumes of information are accessible, the information that can be procured from the information is still should be removed and the extraction is anything but a straight forward issue, rather it presents difficulties.

Each site contains the heft of obstinate content about items, places, motion pictures, and so forth yet the public user can't extricate the conclusions when the information become enormous. Consequently, Opinion mining (OM) frameworks are utilized to separate and sum up required data for the peruser. Opinion Mining has attracted consideration late a long time as a functioning examination territory in arena of the NLP (Natural Language Processing). It is to investigate the following things of peoples’

* Feelings
* Assessments
* Evaluations
* Estimations
* Perspectives
* Attitudes

The above said things are evaluated about the elements. For e.g. items, administrations, associations, people, issues, occasions, themes, and their qualities [1]. Fundamental purpose behind OM being a functioning examination zone is that it consolidates a wide scope of uses in each area. Exploration in OM importantly affects NLP, the executive sciences, political theory, financial aspects, and sociologies as they are totally influenced by peoples’ sentiment.

**1.1. SENTIMENT ANALYSIS**

Feeling investigation is *“characterized as a cycle that systematizes mining of mentalities, feelings, perspectives and suppositions from text, discourse, tweets and information base sources through Natural Language Processing (NLP)”*. Notion investigation comprises assembling sentiments in text into classifications such as "positive" or "impartial" or "negative". The aforementioned furthermore suggested as subjectivity investigation, feeling mining, and evaluation extraction.

The words assessment, opinion, view and conviction are utilized reciprocally however there are contrasts between them.

* Opinion: An end open to question (on the grounds that various specialists have various feelings)
* View: emotional feeling
* Belief: purposeful acknowledgment and scholarly consent
* Sentiment: assessment speaking to one's sentiments

A model for wordings for SA is as given beneath,

<SENTENCE> = Mobile is awesome and camera is very good

<OPINION HOLDER> = <author>

<OBJECT> = <mobile>

<FEATURE> = <camera>

<OPINION >= <awesome><good>

<POLARITY> = <positive>

SA is a task that incorporate numerous undertakings, for example, assumption extraction, supposition characterization, and subjectivity grouping, rundown of assessments or feeling spam location, among others. It plans to examine individuals' slants, perspectives, conclusions feelings, and so on towards components, for example, items, people, subjects ,associations, and administrations.

In term of mathematics, it can speak to an assessment “as a quintuple (o, f, s, h, t), where

o = object;

f = highlight of the article o;

s = direction or extremity of the sentiment on highlight f of item o;

h = assessment holder;

t = time when the assessment is communicated.”

Object: A substance which can be a, individual, occasion, item, association, or point

Highlight: A characteristic (or a piece) of the article regarding which assessment is made. Assessment direction or extremity: The direction of a feeling on a component f speak to either the sentiment is good or neutral or negative.

Assessment holder: The holder of a feeling is the individual or association or a substance that communicates the sentiment.

Lately a great deal of exertion was done in the field of "Assessment Analysis on Twitter" through number of scientists. In the cutting-edge phase, it was anticipated for double characterization which relegates assessments.

**1.2 Twitter Corpora**

Despite the fact that use and ubiquity of Twitter have halted quickly developing and even dropped in late years1, it actually has a lot of faithful clients who continue sharing everything from overall functions to arbitrary individual subtleties with their adherents. It chose to zero in on one of the arbitrary individual subtleties that individuals share, explicitly whatever to sort out with food utilization and connected themes

From the state of art, a few corpora of Indian tweets are exist. Even though, yet nothing of them are space explicit. It have been gathered throughout a broad timeframe. The authors have [1] gathered and investigated one million tweets in India from April 2017 to July 2018 and sixty thousand tweets [2] from November 2016 to March 2017. In another work the authors have [3] gathered and dissected four million tweets of Indian government officials, organizations, media, and clients who collaborated with these elements from August 2016 to July 2018 There are additionally a few informational collections of general conclusion commented on tweets [3-5] adding up to 14,781 tweets altogether.

In this research work, it depicts the Twitter foodie corpus (TFC) and break down its substance.

**1.3 SENTIMENT ANALYSIS APPROACHES**

At hand there are fundamentally two methods for assessment examination for the twitter information:

1.3.1 Machine Learning Approaches

AI based methodology utilizes order strategy to arrange text into classes. There are essentially two sorts of AI methods

* Solo learning:

It doesn't comprise of a class and they don't give the right focuses at all and hence depend on grouping.

* Administered learning:

It depends on named dataset and in this manner the names are given to the model during the cycle. These named dataset are prepared to get significant yields when experienced during dynamic.

Accomplishment of together the learning approaches is predominantly depend on the choice and mining of specific prearrangement of highlights utilized to distinguish supposition. Artificial Intelligence approach appropriate to assumption examination fundamentally has a place through administered characterization. In an machine learning (ML) methods, two preparations of info are required:

1. Preparing dataset

2. Test dataset

Various ML strategies have been figured to arrange the tweets into classes. ML procedures such as

* Maximum Entropy (ME),
* Naive Bayes (NB),
* Support Vector Machines (SVM)

ML begins with collecting training dataset. Next it train a classifier on the preparation information. When an administered arrangement procedure is chosen, a significant choice to make is to choose include. They can disclose to us how reports are spoken to.

The most normally utilized highlights in feeling characterization are

* Term presence and their recurrence
* Part of discourse data
* Negations
* Opinion words and expressions



**Figure 1 Sentiment Classification Based On Emoticons**

Though semi-administered and solo methods are proposed when it is preposterous to expect to have an underlying arrangement of named records/sentiments to group the remainder of things

1.3.2 LBA (Lexicon-Based Approaches)

Vocabulary based technique [17] utilizes feeling word reference with sentiment words and match them with the information to decide extremity. They doles out estimation scores to the feeling words portraying how Positive, Negative and Objective the words contained in the word reference are.

Vocabulary put together methodologies essentially depend with respect to a notion dictionary, i.e., an assortment of known and precompiled supposition terms, states and even sayings, created for conventional classes of correspondence, for example, the Opinion Finder dictionary;



**Figure 2. LBA (Lexicon Based Approach)**

**Tasks of Sentiment Analysis**

Assessment examination is a difficult interdisciplinary undertaking which incorporates characteristic language handling, web mining and AI. It is an unpredictable errand and can be disintegrated into following assignments, via:

* Object/Feature Extraction
* Subjectivity Classification
* Complimentary Tasks
* Sentiment Classification
* Object Holder Extraction



**Figure 3. Tasks of Sentiment Analysis**

**Subjectivity arrangement**

Subjectivity arrangement is the undertaking of ordering sentences as obstinate or not stubborn.

“*Let S = {s1, . . . , sn} be a bunch of sentences in record D. The issue of subjectivity order is to distinguish sentences used to speak to sentiments and different types of subjectivity(subjective sentences set Ss) from sentences used to dispassionately introduce authentic data (target sentences set So), where SsUSo=S”.*

**Estimation Classification**

When the errand of discovering whether a sentence is stubborn is done, it need to discover the extremity of the sentence i.e., regardless of whether it communicates a positive or negative sentiment. Assumption grouping can be a twofold order (positive or negative), multi-class classification (extremely negative, negative, impartial, positive or incredibly positive), regression or positioning.

Contingent on the utilization of estimation investigation, subtasks of assessment holder extraction and article include extraction can be treated as discretionary.

• Opinion Holder Extraction

It is the disclosure of sentiment holders or sources. Identification of assessment holder is to perceive immediate or aberrant wellsprings of sentiment.

• Object/Feature Extraction

It is the disclosure of the objective substance.

**Levels in SA**

Undertakings depicted in the past segment should be possible at a few degrees of granularity.



**Figure 4. Levels of Sentiment Analysis**

**1.4 Organization of this chapter**

This chapter is organized as follows: The section 2 discusses the clarification of the practicality of opinion mining on the mixed languages published in the social media. Section 3 presents the detailed research methodology of this research work. In this chapter, the collected tweets are analyzed with the most esteemed machine learning algorithms. Datasets used for this research work, the experiment setup, the trial results and relating examination is presented detail in the section 4. Section 5 conclude the last area sums up the entire research work and gives an end to this chapter.

1. **LITERATURE REVIEW**

Authors [1] have proposed a model to characterize the tweets as target, positive and negative. They made a twitter corpus by gathering tweets utilizing Twitter API and naturally clarifying those tweets utilizing emojis. Utilizing that corpus, the authors had developed a supposition classifier dependent on the multinomial Naive Bayes strategy that utilizations highlights like N-gram and POS-labels. The preparation set they utilized was less effective since it contains just tweets having emojis. In another work the authors [2] executed two models, a Naive Bayes bigram model and a Maximum Entropy model to arrange tweets. They found that the Naive Bayes classifiers worked obviously superior to the Maximum Entropy model.

In another work, authors [4] planned a two stage programmed feeling investigation strategy for characterizing tweets. They ordered tweets as target or abstract and afterward in second stage, the emotional tweets were delegated positive or negative. The element space utilized included retweets, hashtags, connection, accentuation and shout marks related to highlights like earlier extremity of words and POS. The researchers [5] utilized Twitter streaming information gave by Firehouse API, which gave all messages from each client which are openly accessible progressively. They tested multinomial gullible Bayes, stochastic slope drop, and the Hoeffding tree. They come to an end result that SGD-based model, when utilized with a proper learning rate was the better than the rest utilized.

In another work [6] and [7] utilized Twitter API to gather twitter information. Their preparation information falls in three unique classes (camera, film, portable). The information is named as sure, negative and non-conclusions. Tweets containing sentiments were sifted. Unigram Naive Bayes model was executed and the Naive Bayes rearranging autonomy supposition that was utilized. They additionally disposed of pointless highlights by utilizing the Mutual Information and Chi square element extraction technique. At long last, the direction of a tweet is anticipated for example positive or negative.

The researchers [9] introduced varieties of Naive Bayes classifiers for recognizing extremity of English tweets. Two distinct variations of Naive Bayes classifiers were manufactured to be specific Baseline (prepared to characterize tweets as certain, negative and impartial), and Binary (utilizes an extremity vocabulary and arranges as sure and negative. Unbiased tweets disregarded). The highlights considered by classifiers were Lemmas (things, action words, descriptors and intensifiers), Polarity Lexicons, and Multiword from various sources and Valence Shifters.

From a specialized perspective, the ways to deal with performing OM can be assembled into two classifications, specifically, those of AI based methodology and those of rule based methodology. The AI technique utilizes learning calculations, for example, SVM, NB to decide the conclusion via preparing on a known named dataset, while rule based methodology includes figuring slant extremity for an audit utilizing rules or vocabularies. Research’s [5-7] depend on principle-based methodology while [8 – 14] depend on AI based methodology. All researches recorded here aside from were utilized precision to assess their model. Exactness was determined utilizing the condition of (Correctly anticipated test information/Total number of test information) × 100. Term Frequency (TF) of unigram just as bigram for one of a kind expressions of the film survey corpus were utilized as highlights by shifting estimations of (N - number of tallies). They tried SVM with Linear, Polynomial, Radial Basis Functions, Sigmoid and Precomputed Kernels and got better precision of 64.69% for bigram highlights utilizing Radial Basis Kernel with N=10 than different bits. RKS and LR acquired 61.88% and 57.14% exactness separately utilizing bigrams. MNB and BNB acquired 61.01% and 60.19% exactness individually for unigram highlights.

Seshadiri et al. [10] expected to construct a model to anticipate feelings from Tamil film audits as good or negative. Setting expressions of preparing information, accentuations and punctuation were utilized as highlights. Various methodologies utilizing SVM, Maximum Entropy classifier (Maxent), Decision tree and NB were utilized alongside Tamil SentiWordNet word reference to arrange the audits into positive and negative. SVM acquired an exactness of 75.96% while Decision tree gave a precision of 66.29%. Creators tried the model without utilizing SentiWordNet and acquired 71.91% for SVM and 64.04% for Decision tree. In a research work, the authors [14] have presented the problems and a cost effective methods to mine sentiments from Twitter. Fiercely fluctuating language and spam makes sentiment recovery inside Twitter testing task.

This paper [15] have reviewed the importance of food analysis on social media platform. The researchers [16] have aimed is to find a better food delivery management system between Swiggy and Zomato by performing sentiment analysis on Twitter data. SA is a part of AI that analyzes the subjective information contained in a sentence, which can be an opinion, evaluations, emotions or attitude towards a topic, person or company. In this paper, extracted Twitter data through Twitter API for sentiment analysis on tweets related to Zomato and Swiggy. The proposed research work had calculated the subjectivity and polarity of a sentence to determine whether reviews were positive, negative, or neutral. This calculation helped to decide which food delivery method was better.

1. **RESEARCH METHODOLOGY**

**3.1 Twitter Foodie Corpus (TFC)**

Tweets from 2015 to 2020 have been gathered from the twitter. The twitter foodie corpus is built using the tweets that are collected using the hash tags. They are followed utilizing 362 catchphrases. The phrases are having different emphases of Indian words related with breakfast, eating, lunch, tasting, supper, and so forth. The principle catchphrases are appeared in figure 1. In the collected tweets the words in intense are generally action words that depict eating. From the tweets, these were arched to all serviceable structures then remembered for the complete watchword list. Remainder of watchwords are a bunch of the main sixty food associated arguments which were generally well known in tweets. Apart that it used biryani, dosa, idly, noodles, etc as the extra keywords.



**Figure 5. Tweets keywords**

Toward the start of the undertaking around 16000 drink and food words from gathered tweets were physically explained by means of their individual nominative structures, English interpretations and nutrition classes as indicated by the food control pyramid [7]. Nutritional categories are given in the table 2 below:

**Table 2 List of Nutritional Categories**

|  |  |
| --- | --- |
| **S. No** | **Nutritional Categories** |
|  | Bread, Oats, Rice, Pasta |
|  | Vegetables |
|  | Organic product, Berries |
|  | Milk items |
|  | Meat, Eggs, Fish |
|  | Fats, Oils, Desserts |
|  | Drinks - mixed beverages |
|  | Drinks - non-mixed beverages |

The public delivery incorporates tweet IDs alongside information meadows made inside the extent of this undertaking (beginning through "location\_lng" in figure 7). Total form is accessible based on the singular solicitation intended for research commitments. Corpus additionally incorporates information preparing contents and subtleties on the most proficient method to imitate the analyses.



**Figure 6. Example of a Collected Raw Tweet**

Figure 6 delineates the raw tweet collected to build the TFC in JSON documentation. Individual tweet comprises of essential arenas that are as follows:

* "tweet\_id",
* "tweet\_text",
* "tweet\_author" and
* "created\_at",

The above listed fields will consistently be available, and discretionary fields, which rely upon the metadata and tweet text.

The collected tweet can be separate into 3 gatherings of discretionary fields:

1. "media\_url" and "expanded\_url", it comprises data approximately with the media records as of the tweet
2. "location\_name", "location\_lng", "location\_lat" and "location\_country", it indicates wherever the raw tweet was made
3. "food\_surface\_form", "food\_nominative\_form", "food\_group" and "food\_english\_translation", the tweets are comprise with the semicolon isolated arrangements of nourishments or beverages that show up in a tweet.

***Content Overview***

The in-house built corpus contains 2,00,000 tweets, of which 65841 contain food tweets and 1,29,659 drinks and beverages related tweets. Table 2 and 3 shows the 10 most famous nourishments and beverages from the TFC. Observing from an Indian purchaser viewpoint it is normal that Indians generally vegetables, tea, brew and eat meat, juice, drink water, and natural products. Fascinating, be that as it may, is the high notoriety of desserts, for example, chocolate, cakes, frozen yogurt and Coca-Cola.

**Table – 3 Food related tweets**

|  |  |
| --- | --- |
| Food | Count |
| Biryani  | 11823 |
| Dosa | 8710 |
| Idly | 8657 |
| Noodles | 7513 |
| Fruits | 6361 |
| Egg | 5526 |
| Rice | 4954 |
| Fish  | 4620 |
| Sambar vadai | 4420 |
| Ice cream | 3257 |

**Table – 4 Drink related tweets**

|  |  |
| --- | --- |
| Drinks | Count |
| Water | 64338 |
| Ilaneer | 12104 |
| Coffee  | 19179 |
| Tea  | 16692 |
| Paneer Soda | 1584 |
| Inji Tea | 920 |
| Black Tea | 601 |
| Cocktails | 5766 |
| Pepsi | 4673 |
| Coco-cola | 3802 |

**Figure 7. Food and Drink – Trends in India**

**Figure 8. Food and Drink – Trends in Worldwide**

Figure 7 and 8 shows the annually tally of gathered tweets alongside the expected pattern. It is since for quite a long time 2015 and 2020 just a section has been gathered. The overall notoriety of Instagram and Twitter (a contending informal organization) for our country India which is collected from Google Trends. At hand there was a steady pay of food tweets up until 2017, however from that point forward, it appears to be that the diminishing connects with the general drop in ubiquity of Twitter in India, which is by all accounts straightforwardly inverse to the prevalence of Instagram in India as indicated by Google Trends graph chart.

Figure 7 envisioned the biggest tweet patterns over the previous years from the Indian talking twitter clients. The latest one simply a month back is because the alarm purchasing of buckwheat because of the Covid-19 pandemic of 2020. It is trailed by the multiplying of margarine costs in 2017, India sprat export and boycott Chinese products in 2020 late June.

Figure 8 shows a choice of occasional patterns found the middle value of from information somewhere in the range of 2015 and 2020. Most patterns have one pinnacle zone showing portions of the year once they are extra well known. Instances of this are tea, coffee in December, and pizza, noodles in the late spring. It remained hoping to see dairy milk-silk top high on Valentine's Day. Despite the fact that it tops, thing that matters isn't as greetings.

**3.2 Pre Processing**

The tweets have been initially furnished in the English and other language content with their relating language labels. It performed tokenization or true casing utilizing contents from the Moses Toolkit [10]. It utilized Sentence piece [11] to make a mutual sub-word jargon of 8000 tokens. It supplanted all Twitter-explicit @user makes reference to through “@USR” and “URLs” through “@URL”, as these normally don't comprise applicable etymological information for model to acquire pattern from the datasets. It additionally supplanted various sequential @USR or @URL labels with a solitary one and eliminated them totally on the off chance that they were either toward the beginning of the tweet or at long last. Prior to taking care of the tweets to any preparation stage, they are preprocessed utilizing the accompanying method which are explained below.

***Removal of Noisy Words:***

Usernames (clarified as @username), URLs, and emojis present in the tweets are taken out through and through, while hashtags (commented on as #hashtag) are left all things considered. It likewise tried different things with supplanting emojis by their comparing printed meaning, yet eliminating it prompted for better execution.

***Tokenization:***

Tweets after commotion evacuation are tokenized into sub-words utilizing the word2vec technique and later changed over into their relating IDs.

**Sentiment Annotated Sub-corpus**

It physically explained 5420 tweets stamping them as positive, negative and neutral. The collected tweets are classified using the sentiment classifier and produced 1631 positive, 2507 neutral and 1282 negative tweets. It splits the classified tweets by considering test dataset of 20 percentage tweets and remaining 80 percentage of tweet as a preparation dataset. The pie chart of the labeled dataset is given in the figure 9.

**Figure 9. Labeled Tweets**

**3.3 Machine Learning Algorithms:**

**MULTILAYER PERCEPTRON (MLP)**

A neural organization comprises of handling units, called neurons, where every neuron is associated with different neurons by unidirectional associations of various loads. The yield of each and every preparing unit can be communicated as in (1).

(1)

In the equation 1, where the xj, wj and b are the data sources, loads, predisposition respectively that to neuron and nonlinear initiation work separately. Boundaries of neuron are refreshed as in (2).

(2)

In the equation 2, η is the learning rate. It is the sum where loads are refreshed during the preparation. In any case, it is ambiguous to rough approximation the quantity of the concealed layers and neurons at each concealed layer as they profoundly rely upon the dataset. Extensive number of layers where, neurons will have more boundaries that cannot supportive of any assurance to have better execution. The more the boundaries, the more the examples need in the preparation dataset. The hyper-parameters which will be utilized in the network search to enhance the AUC are as follows:

* Quantity of concealed layers,
* Number of neurons in each concealed layer,
* Enactment work,
* Neuron initializer,
* Cluster size,
* Learning rate,
* Level of dropped neurons,
* Misfortune work,
* Streamlining agent.

***Naive Bayes (NB):***

Naïve Bayes is like the multinomial model. This model is renowned for record order happenings, where paired span event highlights are utilized instead of span frequencies. On the off chance that xi is a Boolean communicating the event or nonattendance of the ith term from the jargon, at that point the probability of a record given a class Ck is given by p(x|Ck) = Πni = 1- pkixi(1-pki)(1-xi) where pki is the likelihood of class Ck creating the term xi. This occasion model is particularly well known for characterizing short messages. It has the advantage of unequivocally displaying the nonappearance of terms. Note that a credulous Bayes classifier with a Bernoulli occasion model isn't equivalent to a multinomial NB classifier with recurrence checks shortened to one. It is a probabilistic classifier and can get familiar with the example of looking at a bunch of reports that has been sorted [9]. It contrasts the substance and the rundown of words to order the records to their correct classification or class. Leave d alone the tweet and C\* be a class that is allocated to d, where



From the above condition, “f” is a ‘feature’, check of highlight (fi) is signified with “ni(d)” and is available in d which speaks to a tweet. Here, m signifies no. of highlights. Boundaries “P(c) and P(f|c)” are processed through greatest probability gauges, and smoothing is used for concealed highlights.

1. **RESULTS AND DISCUSSIONS**

Figure 5 shows a few instances of produced responses to assessment information questions. At hand there were numerous theory responses that were way off the mark to the reference ones yet appeared well and good according to the inquiries, for example, the initial two. There were additionally similarly the same number of or significantly more answers that had neither rhyme nor reason like the last one.

For notion investigation, it performed comparable information pre-preparing ventures concerning question replying, aside from parting words in sentence pieces. It additionally tried different things with lemmatizing and stemming words.

In this section, table 4 demonstrates the consequences of this feeling examination tests. It thought about a Python 3.9 usage of the “Naive Bayes classifier” from NLTK compared to execution of the “Perceptron classifier”. It additionally tried different things with a few mixes of preparing informational indexes - TE (this Twitter Foodie dataset). It found that the most elevated grouping accuracy - 63.47% - is accomplished by utilizing everything for preparing and just stemming all words.

To prepare and group utilizing Naïve Bayes and Multi-Layer Perceptron method, it can utilize the Python NLTK library.

**Table 4 Accuracy of Labeled Tweets**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Stemmed** | **Lemmatized** | **Stemmed + Lemmatized** |
| **Naive Bayes** | 63.21 | 53.32 | 55.72 |
| **Perceptron** | 63.07 | 62.67 | 63.47 |

~~~~

**Figure 10. Accuracy of Labeled Tweets**

1. **CONCLUSION**

In this chapter, the overview of NLP and opinion mining / sentiment analysis is explained with state-of-the-arts. Also, this chapter depicted the production of a genuinely huge corpus related to food and habitats of the user that are available in Twitter. It contributed a few bits of knowledge in generally speaking perceptions picked up from the corpus substance and different patterns that it saw from the information. It accepts that the information would be helpful in numerous etymological, sociological, conduct and other exploration regions.

It explored various ways of developing a food-related inquiry system using a subset of the collected information. As part of this research work, it also explored a concept mining method using a different subset to show possible use cases of this corpus. While the results may not break new ground, it hopes they will inspire relevant future research.

**REFERENCES**

1. D. Milajevs, Language use in a multilingual tweet corpus, in: *Human Language Technologies–The Baltic Perspective: Proceedings of the Eighth International Conference Baltic HLT 2018*, Vol. 307, IOS Press, 2018, p. 88.
2. D. Milajevs, Toward a Comparable Corpus of Indian, Russian and English Tweets, in: *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 26–30. doi:10.18653/v1/W17-2505.
3. M. Pinnis, Indian Tweet Corpus and Investigation of Sentiment Analysis for Indian., in: *Proceedings of the 8th Baltic Conference on Human Language Technologies (Baltic HLT 2018)*, 2018, pp. 112–119.
4. J. Peisenieks and R. Skadin, š, Uses of Machine Translation in the Sentiment Analysis of Tweets (2014).

[5] R. Viksna, Sentiment Analysis in Indian Tweets, Master’s thesis, Rigas Tehniska universitate, 2018. [6] M. Rikters, Universalas metodes Twitter datu analizei, Bachelor’s thesis, Latvijas Universitate, 2012.

1. D. Duston, Food guide pyramid is built on a base of grains, *Daily News* (1992), 8– 8.
2. P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2383–2392.
3. F. Hieber, T. Domhan, M. Denkowski, D. Vilar, A. Sokolov, A. Clifton and M. Post, Sockeye: A Toolkit for Neural Machine Translation, *ArXiv e-prints* (2017).
4. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst, Moses: open source toolkit for statistical machine translation, Association for Computational Linguistics, 2007, pp. 177–180.
5. T. Kudo and J. Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018, pp. 66–71.
6. J.J. Randolph, Free-Marginal Multirater Kappa (multirater *K*free): An Alternative to Fleiss’ FixedMarginal Multirater Kappa, in: *Presented at the Joensuu Learning and Instruction Symposium*, Vol. 2005, 2005.
7. P. Paikens, Lexicon-based morphological analysis of Indian language, in: *Proceedings of the 3rd Baltic Conference on Human Language Technologies (Baltic HLT 2007)*, 2007.
8. S. Bird, E. Loper and E. Klein, *Natural language processing with python O’reilly media Inc*, 2009.
9. Hafiz Muhammad Ahmed, Mazhar Javed Awan, Nabeel Sabir Khan, Awais Yasin, Hafiz Muhammad Faisal Shehzad. Sentiment Analysis of Online Food Reviews using Big Data Analytics. Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 2): pp. 827-836 [http://ilkogretim-online.org/doi:10.17051/ilkonline.2021.02.93](http://ilkogretim-online.org/doi%3A10.17051/ilkonline.2021.02.93)
10. Upadhyay, A., Rai, S., Shukla, S. (2022). Sentiment Analysis of Zomato and Swiggy Food Delivery Management System. In: Luhach, A.K., Poonia, R.C., Gao, XZ., Singh Jat, D. (eds) Second International Conference on Sustainable Technologies for Computational Intelligence. Advances in Intelligent Systems and Computing, vol 1235. Springer, Singapore. <https://doi.org/10.1007/978-981-16-4641-6_4>
11. Soussan, T., Trovati, M. (2022). Sentiment Analysis: A General Review and Comparison. In: Barolli, L., Miwa, H. (eds) Advances in Intelligent Networking and Collaborative Systems. INCoS 2022. Lecture Notes in Networks and Systems, vol 527. Springer, Cham. <https://doi.org/10.1007/978-3-031-14627-5_22>