Informatics Bulletin, Faculty of Computers and Artificial Intelligence, Helwan University
Published Online *Vol 3 Issue 3, October 2021*
(https://fcihib.journals.ekb.eg)

FCI-H
Informatics Bulletin

A survey on places clustering based on sentiment analysis

Eslam Hanafy, Hala Abdelgelil, Soha Ahmed Ehssan
Computer Science Department - Faculty of Computer and Artificial Intelligence - Helwan University - Cairo, Egypt
islamhanafy1@hotmail.com, aehala2000@yahoo.com, dr.soha@fci.helwan.edu.eg

***Abstract*— Sentiment analysis is an automated technique for extracting opinions, emotions, and sentiments from text and records internet-based attitudes and feelings. Individuals offer their perspectives on various topics through blog entries, comments, reviews, and tweets. Opinion mining and sentiment analysis can be used to monitor places, products, and brands and then determine if they are viewed positively or negatively. . Classification and clustering techniques are generally used for sentiment analysis. But, clustering-based techniques are compelling for sentiment analysis from the text. Although their results are subject to alteration depending on the data pre-processing method used or the terms weighting approach used, they have a significant advantage over supervised learning methods. Furthermore, clustering-based techniques can produce satisfactory results without needing organized data, linguistic knowledge, or training time. This paper will review recent works in sentiment analysis techniques and places' clustering techniques based on sentiment analysis in aim to cluster places based on safety measures during corona time for better user satisfaction.**

***Index Terms*— Classification, Clustering, places clustering, Clustering ensembles, Consensus Clustering, sentiment analysis, Support Vector Machine, Random Forest.**

I. Introduction

Social media has evolved into a primary activity in our everyday lives. According to Statista[29], the average person spends 145 minutes per day on social media networks. People are no longer merely using social media platforms for conversing; instead, it has evolved into their virtual reality. People connect with new acquaintances and make purchases through them, and they even offer reviews on everything on social media. We can now agree that review and suggestion posts changed numerous enterprises and swung numerous thoughts and emotions, affecting our lives on a social, political, and personal level. Now, as a result of the pandemic, people's criteria of ranking have shifted. People are now more concerned with the establishment's dedication to safety and pandemic avoidance guidelines than with the previous ranking requirements. Previously, regulations included food quality in restaurants and modes of entertainment in clubs, but since the new covid pandemic, safety standards have evolved to include social distance commitment, cleaning locations and equipment, and wearing a face mask.

Our primary objective is to create a framework that will assist in ensuring that places comply with pandemic avoidance safety standards and cluster places according to safety measures. We will evaluate previously employed top-performing sentiment analysis and place clustering techniques and combinations in order to achieve the best outcomes in our subsequent work. The rest of the paper is structured as follows. Section 2 briefly covered the dataset pre-processing techniques. Section 3 discussed The vectorization techniques. Sections 4,5 introduced, summarized, and compared several related papers to sentiment analysis and clustering techniques and their objectives and findings. Finally, section 6 concluded the paper and discussed the upcoming future work.

II. Dataset Pre-processing

Data can be downloaded from different repositories like Kaggle, Data world, UCI, ASU, Open Knowledge Labs, Kdnuggets, SNAP

In addition, data can be collected directly from Twitter using:

1- Twitter APIs:
   a. search API: collect the user Twitter data based on the hashtags.
   b. Stream API: get the real-time data from Twitter.
2- Tools to Download Twitter Data: Twitter's official archivedownload, BirdSong Analytics, Cyfe, NodeXL, and TWChat

Also, data can be crawled from other websites using web crawlers and spiders, and other APIs that are available online for data collection

Alternatively, manual data collection if needed.

A. Lower casing

Lower-casing is a simple, effective standard text pre-processing technique that converts all the characters of the text into lower case format.

Lower-casing is more helpful for text featurization techniques like frequency, TF-IDF. it helps to combine the exact words, thereby reducing the duplication and getting correct counts / TF-IDF values.[1]

Lower-casing may be ineffective for specific tasks, such as

Part of Speech tagging (in which proper casing provides information about Nouns and so on) and Sentiment Analysis (where upper case refers to anger and so on).

B. Punctuations Removal

The punctuation removal technique used to remove the punctuations from the text for example these symbols “ (!"#$%&\'()*+,-./:;<=>?@[\\]^_{|}~`) ” to standardize the text.

C. Tokenization

Tokenization is decomposing a character sequence into tokens (words/phrases) and removing unnecessary characters such as punctuation marks. Following that, the list of tokens is used for further processing.[2]

D. Stop word Removal

Stop words are a group of words that are frequently employed in a language. They are necessary for human reading in order to comprehend the language rapidly. In English, stop words include "a", "the", "is", "are", and "and". However, these terms are not as useful for machine-reading as they are for human reading; thus, by deleting common information words from the text, we can concentrate on the important words instead, which improves the system’s size and performance.[3]

E. Convert Accented Characters

Many languages have accent marks in some languages like Arabic, French, Spanish, Portuguese, and german are using accent marks repeatedly, even English have some commonly used accent marks, for example, acute accent (́) as in *Café*, grave accent (`) as in *Cortège*, diaeresis mark (¨) as in *Naïve, and many other symbols and words. Humans can recognize that it is the same word* with or without the accent mark, but for machines we the word has to be converted and standardized like “café” and “cafe” to just “cafe” so that the model does not treat them as different words even though they are referring to the same thing.[4]

F. Treatment for Numbers

The treatment of numbers consists of two steps.
The first step is to convert numbers from their word form to their numeric form in order to standardize the language. [4]
The second stage is to eliminate superfluous numbers to the proposed model; nevertheless, we will leave them in place if required.

G. Lemmatization

Lemmatization is the task that takes into account the morphological analysis of the words, that is, collecting the numerous inflected forms of a word into a single form for analysis. In other words, lemmatization techniques attempt to map verb forms to an infinite number of tenses and noun forms to a single form. [5]

H. Stemming

Stemming It streamlines the sentiment analysis process by reducing words to their stem or root form. The same word might take on a distinct connotation depending on the grammatical context, as in “contains, contained, containing.” stemming restored it to its infinitive state, which is “contain.” [6]

III. VECTORIZATION

To allow for more formal descriptions of the algorithms, we first define some terms and variables that will be frequently used in the following: Given a collection of documents D = {d1, d2, ..., dD}, let V = {w1, w2, ..., wv} be the set of distinct words/terms in the collection. Then V is called the vocabulary. The frequency of the term w ∈ V in document d ∈ D is shown by fd (w), and the number of documents having the word w is represented by fD (w). The term vector for document d is denoted by (t_d) → = (fd (w1), fd (w2), . . ., fd (wv)) [5].

The most common way to represent documents is to convert them into numeric vectors. This representation is called the “Vector Space Model” (VSM). Even though its structure is simple and originally introduced for indexing and information retrieval [7], VSM is broadly used in various text mining algorithms and IR systems and enables efficient analysis of an extensive collection of documents [8].

In VSM, each word is represented by a variable having a numeric value indicating the weight (importance) of the word in the document. There are two primary term weight models:

1- Boolean model: In this model, a weight ωij > 0 is assigned to each term wi ∈ dj. For any term that does not appear in dj,ωij = 0.
2- Term frequency-inverse document frequency (TF-IDF): The most popular term weighting scheme is TF-IDF. Let q be this term weighting scheme, then the weight of each word w ∈ d is computed as follows:

$$q(w) = f_d(w) * \log \frac{|D|}{f_D(w)}$$

Where |D| is the number of documents in the collection D. In TF-IDF, the term frequency is normalized by inverse document frequency, IDF. This normalization decreases the weight of the terms occurring more frequently in the document collection, making sure that the matching of documents is more affected by uncommon words with relatively low frequencies.

Based on the term weighting scheme, each document is represented by a vector of term weights ω(d) = (ω (d, w1), ω (d, w2),..., ω (d, wv)).[17][18] We can compute the similarity between two documents d1 and d2. One of the most widely used similarity measures is cosine similarity and is computed as follows:

$$s(d_1, d_2) = \cos(\theta) = \frac{d_1 . d_2}{\sqrt{\sum_{i=1}^{v} {w_{1i}}^2} . \sqrt{\sum_{i=1}^{v} {w_{2i}}^2}}$$

IV. PLACE ANALYSIS AND CLUSTERING

Luiz et al. developed a framework for combining classifiers and clusters in order to produce a more consolidated classification. They used the C3E-SL algorithm to combine the support vector machine (SVM) classification technique with Clustering Ensembles. They employed a variety of datasets to demonstrate the algorithm’s efficacy at improving the outcomes. They used four different Twitter datasets (Health care reform (HCR) 1286 tweets, Obama-McCain Debate (OMD) 1906 tweets, Sanders-Twitter Sentiment Corpus 1224 tweets, Stanford-Twitter Sentiment Corpus 1.6 million tweets). The Combined methods resulted in a significant increase in the analysis's accuracy compared to utilizing SVM alone. [10]

Muhammad Afzaal et al. proposed a framework that extracts information about tourism from Twitter. They analyzed three types of opinion mining trend-based, aspect-based, and sentence-based. They classified the tweets by identifying the aspects and opinion trends. They used Twitter data using search API (British museum 2000 tweet, Tower of London 2000 tweet, London eye 2000 tweet). They could quickly extract information and visualize it for the user, providing valuable and meaningful information about the place the user wishes to visit without reading the reviews. [11]

Venu Dave et al. proposed a sentiment analysis-based model for processing reviews of touristic places collected from a website. They gathered reviews from a variety of sources, including TripAdvisor and goibibo.com. They constructed a term-document matrix from the dataset and then used a naïve Bayes classifier to classify the terms. They achieved an accuracy of 83.00% using naïve Bayes to classify sentiment based on four emotions (joy, surprise, anger, and unknown). [12]

Sumbal Riaz et al. conducted phrase-level sentiment analysis on customer reviews. They used web crawlers and spiders to acquire 1.2 million reviews for 20,821 products across six major categories from 0.9 million reviewers from Amazon and various online shopping websites. They employed TF-IDF vectorization in conjunction with the k-means clustering technique to place the words in various clusters according to their intensity. They calculated the intensity of sentiment polarity by measuring its strength. Additionally, they compared their technique's results to the star ratings of the same data and discovered a more positive and neutral feeling toward products. [13]

Muhammad Afzaal et al. proposed an aspect-based sentiment classification method for tourist reviews in order to identify aspects and classify the sentiments. They gathered reviews via crawlers and APIs from social media networks. The city of interest was decided to be London. They gathered 2000 reviews in the restaurant category and 4000 in the hotel category. They examined various classifiers for aspect-based classification and discovered that the Naïve Bayes multinomial classifier had the highest performance on the dataset they used. They attained an accuracy of 88.08 % on the restaurant dataset and 90.53 % on the hotel dataset. [14]

Apeksha Arun Wadhe et al. proposed a model for touristic places classification based on tourists' reviews. They Collected the dataset from tourists' reviews variety of tourism-related websites. They tested various combinations of feature extraction methods (Count Vectorization, TF-IDF) and classifiers (Naïve Bayes, Support Vector Machine, Random Forest). They obtained the best results by combining Term Frequency Inverse Document Frequency with Random Forest (TF-IDF + RF). They obtained an accuracy of 86.13 %. They opted for more accuracy over time. [15]

Wen Chen, Zhiyun Xu et al. proposed a classification system based on tourists' reviews. They used Python Crawler to obtain reviews from trip advisors (Mutianyu Great Wall 2772 review, Wizarding World of Harry Potter 6641 review, Tower of London 4428 review, Sydney Opera House 6776 review). They used Word2vec-based text vectorization and knowledge graph-based keyword semantic expansion method sampling with support vector machine classification. They achieved an accuracy of 82.3 % in Harry Potter's Wizarding World, 84.3 % at the Sydney Opera House, and 93.3 % at the Tower of London. [16]

Gang Li, Fei Liu, proposed new techniques to extend the capability of clustering in two aspects: first, by applying opposite opinion contents processing and non-opinion contents processing techniques to enhance accuracy further, and second, by using a modified voting mechanism and distance measurement method to conduct fine-grained (three classes) sentiment analysis. They produced higher accuracy in binary sentiment analysis by applying opposite opinion contents and non-opinion contents processing. [27]

Sisi Liu et al. introduced a systematic framework for sentiment clustering using topic and temporal features for large Email datasets. They collected 32,716 Email messages from the Enron corpus database. They employed the bag-of-words model as a term weighting approach and revised the DBSCAN algorithm for clustering. They were able to verify the legitimacy of emails through the time they were being sent through. they validated the parameters chosen for DBSCAN by comparing items in different clusters and the same cluster. [28]

Korawit Orkpholm et al. proposed a combined sentiment analysis-based technique for clustering similar tweets together. They collected 1000 nonredundant tweets using the Twitter API by specifying a keyword (iPhone x). They used the term frequency-inverse document frequency (TF-IDF) to vectorize the messages, and used singular value decomposition (SVD) to reduce dimensions by selecting the most critical relevant terms for the K-means clustering technique, and used artificial bee colony (ABC) to optimize k-means by finding the best initial state of centroids this produced so much better accuracy than the normal k-means (around 41% improvement over normal K-means using the silhouette metric). They achieved a final overall average accuracy of 75 %. [25]

Baojun Ma et al. conducted a thorough experiment to demonstrate the effect of data pre-processing, term weighting models, and clustering algorithms on online review sentiment clustering performance. They used datasets from a variety of online sources. They experimented with various feature extraction and clustering techniques. They discovered that whereas K- means clustering algorithms exhibit apparent advantages on balanced review datasets while performing rather poorly on unbalanced datasets by considering clustering accuracy. Meanwhile, some clustering algorithms such as spectral clustering showing relatively poor performances on balanced datasets may perform very well on the unbalanced ones. The newly designed weighting models (BM25, DPH-DFR, and H-LM) are somewhat better than the traditional weighting models (Binary, TF, and TF-IDF) for sentiment clustering on both balanced and unbalanced datasets. Adjective and adverb word extraction strategies improve clustering performance while adopting stemming and stop-word removal strategies would negatively influence sentiment clustering. [26]

Objective	Method / algorithm	dataset	Pros	Cons
Enhance sentiment analysis through combining classification and clustering techniques [10]	Combined support vector machine with Consensus between Classification and Clustering Ensembles based on Squared Loss SVM +C^{3}E-SL	4 different twitter datasets 1- Health care reform (HCR) 1286 tweet 2- Obama-McCain Debate (OMD) 1906 tweets 3- Sanders - Twitter Sentiment Corpus 1224 tweets 4- Stanford - Twitter Sentiment Corpus 1.6 million tweets	C3E-SL algorithm enhanced the results of The Support vector machine	They used small datasets + only tried SVM classifier with one type of ensemble clustering and this made some ambiguity
Classification of the extracted information on the basis of tourist places aspects [11]	POS tagger + aspect-based opinion mining	Twitter data using search API 1- British museum 2000 tweet 2- Tower of London 2000 tweet 3- London eye 2000 tweet	they were able to use aspect classification and trend classification to get useful information about the places	Important aspects may not be extracted if they are infrequent or implicit.
Process the reviews of tourist places, collected from a website based on sentiment analysis [12]	Naïve Bayes Classifier	Review from different websites including trip-advisor and goibibo.com	They were able to enhance the accuracy to 83 % on the same dataset	They failed to choose the right classes because as example surprise could be good or bad class
Cluster products' reviews based on sentiment polarity [13]	TF-IDF + k-means clustering + Euclidean distance calculation	Web crawlers and spiders from amazon and various online shopping websites. 1.2 million reviews, of product belong to 6 major categories, by 0.9 million reviewers for 20,821 products.	The TF-IDF enhanced keyword extraction, they achieved better results than star ranking techniques	They categorized majority of reviews as neutral
Aspect-based sentiment classification for tourist reviews [14]	Naïve bayes multinomial	Reviews from social media websites using crawler and APIs London was chosen as the city of interest 1- restaurant domain 2000 review 2- hotel domain 4000 review	They compared different techniques for classification and compared results considering accuracy and complexity and achieved high accuracy	
Tourist place classification based on sentiment analysis [15]	Term Frequency -Inverse Document Frequency Vectorization + Random Forest	Collected dataset from many tourism websites in CSV format	They compared different combinations of feature extraction algorithms and classification techniques	They could use different feature selection method like Recursive feature elimination with cross-validation to improve the accuracy
Sentiment analysis and classification from tourists reviews for tourism decision making [16]	Support Vector Machine + Random Over Sampler + Word2vec + Knowledge Graph	Python Crawler to obtain reviews from trip advisor 1- mutianyu Great Wall 2772 review 2- Wizarding World of Harry Potter 6641 review 3- Tower of London 4428 review 4- Sydney Opera House 6776 review	combining (Random Over Sampler + Word2vec + Knowledge Graph) sampling techniques enhanced the classification greatly	They could try some other classification techniques to enhance the accuracy
Clustering microblogs based on sentiment analysis [25]	term frequency–inverse document frequency (TF–IDF) + singular value decomposition (SVD) + K-means with artificial bee colony (ABC)	1000 non redundant tweet through twitter API	They combined new algorithms with k-means clustering which opened the way for new research points	They could use better version of ABC, they consumed much more time and resources (65 times than the normal K-means)
Exploring clustering techniques and weighting models and Comparing their combinations accuracy on document sentiment analysis [26]		8 different datasets from different places (amazon – trip advisor) and other websites	they tried new weighting models (Okapi BM25, Hierarchal Linear Model, different hyper-geometric Divergence from Randomness) and proved their advantage over the old models (e.g., Binary, TF and TF_IDF)	
Develops a systematic scheme of approach for discovering sentiment distribution patterns from large Email corpus [28]	bag-of-words model as term weighting approach and revised DBSCAN algorithm for clustering	32,716 Email messages extracted from the Enron corpus database	they clustered large corpus and used relatively large dataset	They didn't try to compare with other clustering techniques

VI. CONCLUSION

In this paper, the most recent techniques for clustering were reviewed based on sentiment analysis. We discussed the most often used methods for typical tasks like data collection, pre-processing, vectorization techniques, analysis, and clustering techniques. Additionally, several studies were presented that explores novel ways to combine classification and clustering approaches or enhance existing clustering techniques through new algorithms or vectorization techniques. Additionally, a comparison between reviewed papers was added in an attempt to demonstrate the advantages and disadvantages of each work from our perspective. In the future; a place clustering model that complies with the new pandemic safety standards will be proposed, as we could not discover any earlier work that covered the same idea. We will work on a more prominent and recent dataset as there is a need to collect data that can help us to model the place's safety. Experiments will be done using the techniques discussed in this paper. Additionally, we can try additional combinations of techniques as well.

REFERENCES

[1] https://www.kaggle.com/sudalairajkumar/getting-started-with-text-preprocessing

[2] Jonathan J Webster and Chunyu Kit. 1992. Tokenization as the initial phase in NLP. In Proceedings of the 14th conference on Computational Linguistics-Volume 4. Association for Computational Linguistics, 1106–1110

[3] https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html

[4] https://towardsdatascience.com/nlp-text-preprocessing-a-practical-guide-and-template-d80874676e79

[5] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", arXiv:1707.02919v2 [cs.CL] 28 Jul 2017

[6] Khan, F.H., Bashir, S., Qamar, U.: TOM: Twitter opinion mining framework using hybrid classification scheme. Decis. Support Syst. 57, 245–257 (2014)

[7] Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. Commun. ACM 18, 11 (1975), 613–620

[8] Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß. 2005. A Brief Survey of Text Mining. In Ldv Forum, Vol. 20. 19–62

[9] https://www.geeksforgeeks.org/web-information-retrieval-vector-space-model/

[10] Coletta, Luiz, Felix, Nadia, Hruschka, Eduardo, Hruschka, Estevam, "Combining Classification and Clustering for Tweet Sentiment Analysis," 2014, DO - 10.13140/2.1.1060.9281

[11] Muhammad Afzaal, Muhammad Usman, "A Novel Framework for Aspect-based Opinion Classification for Tourist Places," 2015

[12] Venu Dave, Dhwani Shah, DikshiS uthar, Bhagirath Prajapati, Priyanka Puvar, "Sentiment Analysis of Tourists Opinions of Amusement, Historical and Pilgrimage Places: A Machine Learning Approach", 2017, DO - 10.14445/22312803/IJCTT-V46P111

[13] Sumbal Riaz1, Mehvish Fatima1, M. Kamran1, M. Wasif Nisar, "Opinion mining on large scale data using sentiment analysis and k-means clustering", 2017

[14] Muhammad Afzaal, Muhammad Usman, Alvis Fong, "Tourism Mobile App with Aspect-Based Sentiment Classification Framework for Tourist Reviews", 2019 DOI 10.1109/TCE.2019.2908944, IEEE

[15] Apeksha Arun Wadhe, Shraddha S. Suratkar, "Tourist Place Reviews Sentiment Classification Using Machine Learning Techniques", 2020

[16] Wen Chen, Zhiyun Xu, Xiaoyao Zheng, Qingying Yu, Yonglong Luo, "Research on Sentiment Classification of Online Travel Review Text", Appl. Sci. 2020, 10, 5275; doi:10.3390/app10155275

[17] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. Information processing & management 24, 5 (1988),513–523

[18] Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. Commun. ACM 18, 11 (1975), 613–620.

[19] https://www.upgrad.com/blog/clustering-vs-classification/

[20] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[21] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naive Bayes for text categorization revisited," in Australasian Joint Conf. Artificial Intell., 2004, pp. 488-499, DOI:10.1007/978-3-540-30549-1_43

[22] https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/

[23] https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[24] Peter D. Caie BSc, MRes, PhD, ... Ognjen Arandjelović M.Eng. (Oxon), PhD (Cantab), "Precision medicine in digital pathology via image analysis and machine learning", 2021

[25] Korawit Orkpholm, Wu Yang, "Sentiment Analysis on Microblogging with K-Means Clustering and Arti¯cial Bee Colony", 2019, DOI: 10.1142/S1469026819500172.

[26] Baojun Ma, Hua Yuan, Ye Wu "Exploring performance of clustering methods on document sentiment analysis", 2015, Journal of Information Science 2017, Vol. 43(1) 54–74.

[27] Gang Li, Fei Liu, "Sentiment analysis based on clustering: a framework in improving accuracy and recognizing neutral opinions", 2013, DOI 10.1007/s10489-013-0463-3

[28] Sisi Liu, Guochen Cai, Ickjai Lee, "Sentiment Clustering with Topic and Temporal Information from Large Email Dataset", 2016

[29] https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/