A survey of artificial intelligence techniques for user perceptions' extraction from social media data

Abstract. Measuring and analyzing user perceptions and behaviors in order to make user-centric decisions has been a topic of research for a long time even before the invention of social media platforms. In the past, the main approaches for measuring user perceptions were conducting surveys, interviewing experts and collecting data through questionnaires. But the main challenge with these methods was that the extracted perceptions were only able to represent a small group of people and not whole public. This challenge was resolved when social media platforms like Twitter and Facebook were introduced and users started to share their perceptions about any product, topic, event using these platforms. As these platforms became popular, the amount of data being shared on these platforms started to grow exponentially and this growth led to another challenge of analyzing this huge amount of data to understand or measure user perceptions. Computational techniques are used to address the challenge. This paper briefly describes the artificial intelligence (AI) techniques, which is one of the types of computational techniques available for analyzing social media data. Along with brief information about the AI techniques, this paper also shows state-of-the-art studies which utilize the AI techniques for measuring user perceptions from the social media data.

Keywords: social data analysis, social media data, NLP, machine learning, deep learning, transfer learning, user perceptions, sentiment analysis, topic modelling, perception extraction, user perceptions.

1 Introductions

User perceptions about any events, topics, policies, etc. have always appealed the attention of policy and decision makers. These perceptions are always considered as strong evidence for making and adjusting user-centric decisions [1-3]. The traditional method of analyzing/investigating user perceptions is usually based on data collection from survey polls and questionnaires. Next, the collected data is analyzed using traditional qualitative and quantitative methods [4, 5]. However, some of the researchers have argued that these approaches more likely represent a small group of individual user perceptions rather than public user perceptions [6, 7]. Furthermore, due to the time and cost constraints involved in survey and questionnaire activities, the amount of collected data is very limited and hence it restricts the overall findings for understanding user perceptions to a large extent [8].

Nowadays, social media platforms like Twitter, Facebook, LinkedIn, etc. offer a new way of understanding and measuring user perceptions. There has been an increase in adoption and use of social media platforms by the general public as well as the enterprises, industry owners, government officials, scientists, scholars, etc. [9]. The understanding and extraction of user perceptions from social media data has been widely studied in several domains including social science [10], education [11], politics [12], marketing [13], healthcare [14], finance [15] and disaster management [16]. Hence, using social media data as a data source for user perception extraction and analysis can overcome the limitations of traditional surveys and questionnaires methods [17]. It is helpful in forecasting future user perceptions related to any event, topic or policies.[18].

Xuefan et al. in [19] conducted a review on perception extraction and understanding on social media data. According to the review results, Twitter is the most widely used social media platform for perceptions extraction along with Facebook and other platforms. The most frequent keywords of the studies included in the review are social media, twitter, sentiment analysis, public perception, public engagement, opinion mining, NLP, and perceptions. The major techniques used for perception extraction from the studies included in the review are sentiment analysis and topic modelling. Table 1 shows some of the recent studies where perception extraction techniques from social media data are used for several domains like health, business, tourism, etc.

Also,
Fig. **1** shows the summary of state-of-the-art studies mentioned in Table 1 for perception extraction from social media data for several domains. As we can see from
Fig. **1** that the maximum number of studies were conducted in the year 2019 with Twitter as the most selected social media platform. Also, the studies belong to several domains like tourism, business, health, etc.

The rest of the paper is organized as: Section 2 briefly describes the different AI techniques including machine learning, deep learning and natural language processing for social media data analysis. Section 3 discusses the state-of-the-art studies available for social media analysis using the techniques explained in section 2. Section 4 shows the existing challenges in the state-of-the-art studies. Section 5 concludes the overall work.

Table 1. State-of-the-art studies for perception extraction from social media data

#	Title	Year	Domain	Social Media Platform
1	Social media data analysis to predict mental state of users using machine learning techniques [20]	2021	Health	Twitter
2	The application of artificial intelligence and data integration in COVID-19 studies: a scoping review [21]	2021	Health, COVID-19	Twitter
3	Exploring temporal suicidal behaviour patterns on social media: Insight from Twitter analytics [22]	2020	Health	Twitter
4	Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of twitter data [23]	2020	Health	Twitter
5	Analyzing social media, analyzing the social? A methodological discussion about the demoscopic and predictive potential of social media [24]	2020	Politics	Twitter
6	A case study in belief surveillance, sentiment analysis, and identification of informational targets for e-cigarettes interventions [25]	2019	Health	Twitter
7	Realizing social-media-based analytics for smart agriculture [26]	2019	Agriculture	Twitter
8	Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor [27]	2019	Tourism	TripAdvisor
9	Leveraging social media to gain insights into service delivery: a study on Airbnb [28]	2019	Tourism	AirBnB Reviews
10	Using Classification Technique for Customer Relationship Management based on Thai Social Media Data [29]	2019	Business	Facebook
11	A new approach of social media analytics to	2019	Business	Twitter
12	predict service quality: evidence from the airline industry [30]			
13	Topic modeling and sentiment analysis of global climate change tweets [31]	2019	Climate Change	Twitter
14	Identifying racist social media comments in Sinhala language using text analytics models with machine learning [32]	2018	Social Issues	Facebook
15	The Twitter Bullishness Index: A Social Media Analytics Indicator for the Stock Market [33]	2016	Stock Market	Twitter
16	Analyzing Twitter to explore perceptions of Asian restaurants [34]	2016	Tourism	Twitter

Fig. 1. Summary of perception extraction techniques from social media platforms

2 Background

The fast-growing use of social media platforms and their relevant application areas have made major advancements in the different ways which people use to interact with each other [35]. The in-depth analysis of social media data is sometimes difficult because of its unpredictable nature due to several facts like the data is dynamic, wide-ranging, and scattered. The recent advances in computational techniques/methods like artificial intelligence (AI) have made in-depth analysis of social media data quite easier. These techniques help in understanding several patterns on social media platforms like social media usage, online behaviour, data/content sharing, perceptions of different types of people about certain topics, etc. [36]. The extraction of these patterns can give a variety of benefits to organizations, governments, and non-profit organizations to design their services and policies focusing on user-centric methodology [37]. There have been a lot of attempts in literature for extracting valuable insights from vast social media data for better decision making. Some of the examples of such insights are analyzing opinions of users towards different products, analyzing election results, and understanding users' behaviour [32-34]. This section will further discuss different types of AI techniques being used for analyzing social media data.

The AI techniques/methods are all about "making machines intelligent" by using a variety of approaches. The field of AI has been around almost more than six decades, and it has faced many ups and downs throughout this period. In the starting days, AI research showed a lot of promises to the communities, but those promises were somehow not fulfilled due to unavailability of digital data and computational power. This was the period of late 1980s and early 1990s and termed as "AI Winter", where not much progress was achieved in terms of solving real-world problems using AI. However, later, when new techniques were introduced along with the availability of digital data and huge computational powers, AI started to increase in popularity again. Here, digital data is mainly referred to online web and social media platforms data. The different AI techniques/methods for analyzing social media data both (structured and unstructured) can be divided into three major types based on use case and ultimate goal to be achieved. Fig. 2 shows the three different types of AI techniques/methods for analyzing social media data to understand user behaviors, usage patterns, communications and perceptions.

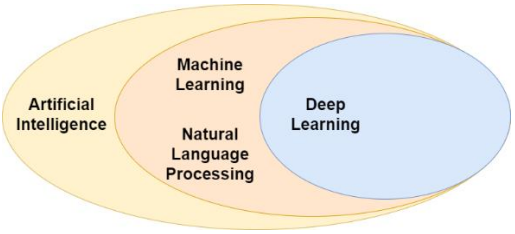


Fig. 2. Types of AI techniques

1.1 Machine Learning (ML)

Machine learning (ML) is a type of AI technique, used to automate solutions for complex problems that are very difficult to solve using general hand-crafted rules-based approach. This technique does not require any explicit rules/steps to design the solution, instead it learns different set of rules/steps from the set of provided data relevant to the real-world problem that needs to be solved. For example, in case of handwritten character recognition from images, the data can be the collection of several images with variety of numerical characters written by different set of people. In other words, this technique learns patterns, relations from the data and the larger the data, the better ML learns [38-40]. Since ML considers using all the data provided for learning the rules, it becomes more accurate as compared to hand-crafted rules because there is no human-bias involved while defining the rules. As ML is a type of automatic learning from the data to solve any problem, the learning is categorized into three different methods shown in Fig. 3. These learning methods enable ML to learn from the available data. The data is generally of two types [41]:

- **Labelled Data**: The data available with its relevant answers/labels. For example, collection of raw images of dogs and cats along with labels that specify which images are dogs and which are cats, respectively.
- **Unlabeled Data**: The data available without the relevant answers/labels. For example, collection of raw images of dogs and cats but without any labels provided per image.
  It is also worth to discuss here the concept of training, validation and test data before going into further discussions of learning methods.
- **Training Data**: The data used to train any ML model so that the model can learn the patterns and behaviors inside/from the data.
- **Validation Data**: The data used during the training process to assess the performance of the ML model during each training step.
- **Test Data**: The new unseen data used to measure the performance of final trained ML model to evaluate its learning/capability to make decisions.

Fig. 3 shows the detailed taxonomy of ML and its learning methods. The taxonomy is further discussed below.

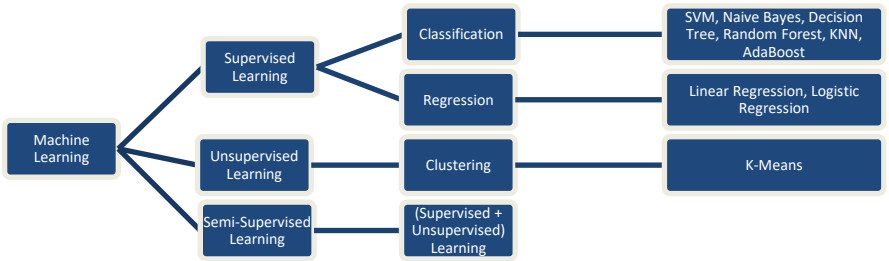


Fig. 3. Taxonomy of ML

1.1.1 Supervised Learning

In supervised learning method, the ML model is given a collection of data specific to any real-world problem along with answers/labels (i.e. labelled data). The provided data is in the form of mapping from *Input(x)* → *Output(y)*, where x represents the list of data points and y represents the corresponding answers/labels. The ML model has to learn different relationships/rules of data points to their corresponding labels. For example, a dataset/collection of emails with provided labels as spam or non-spam is a type of labelled data and the learned ML model that can identify any new email as spam or non-spam is a type of supervised learning method. Furthermore, the supervised learning methods are divided into two types: Classification [42] and Regression [43].

1.1.2 Unsupervised Learning

In unsupervised learning method, the ML model is given a collection of data specific to any real-world problem without any answers/labels (i.e., unlabeled data). The provided data is in the form of *Input(x)* only, where x represents the list of data points without any labels. The ML model has to identify and differentiate different data points and distribute these into different groups based on its learning data. For example, a dataset/collection of shopping patterns from several customers. The unsupervised model learns to group the customers based on buying patterns. This type of unsupervised learning is commonly called Clustering [44].

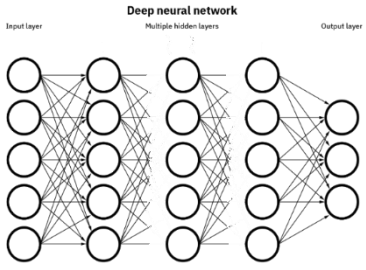
1.1.3 Semi-Supervised Learning

The semi-supervised learning model is a combination of both supervised and unsupervised learning as supervised learning requires a lot of labelled data and this labelling requires a lot of time and human effort. Therefore, the semi-supervised learning model learns from the small amount of labelled data in supervised manner and then uses the unsupervised learning to label rest of the remaining data points [45]. Furthermore, more detailed information regarding ML, its learning methods and models is also discussed in [46].

1.2 Deep Learning (DL)

Deep learning is a sub-type of ML which mimics the same way that humans use to gain or understand certain types of information and knowledge. The major difference between ML and DL is the composition of its models where ML models are linear in nature and DL models are stacked hierarchical and complex in nature [47]. Another advantage of using deep learning is the automatic learning of important features from the raw data contributing towards decision making. The machine learning that we discussed previously is more often dependent on human intervention to learn. Deep learning is also referred as “Deep Neural Networks (DNNs)”. The simple neural networks, also referred as “Artificial Neural Networks (ANN)” are structured and motivated by human brain reflecting the same way the biological neurons work [48]. The ANNs are consist of hierarchical node layers containing an input later, one or more than one hidden layer and an output layer. The DNNs are the special kind of ANNs with more than one hidden layer.

ig. 4 shows the basic architecture of a simple DNN.



ig. 4. Basic Architecture of a Simple DNN [1]

Like ML, there are several types of DNNs based on the objective/problem that needs to be achieved/solved using deep learning. Fig. 5 shows the detailed taxonomy of DL/DNNs and is further discussed below.

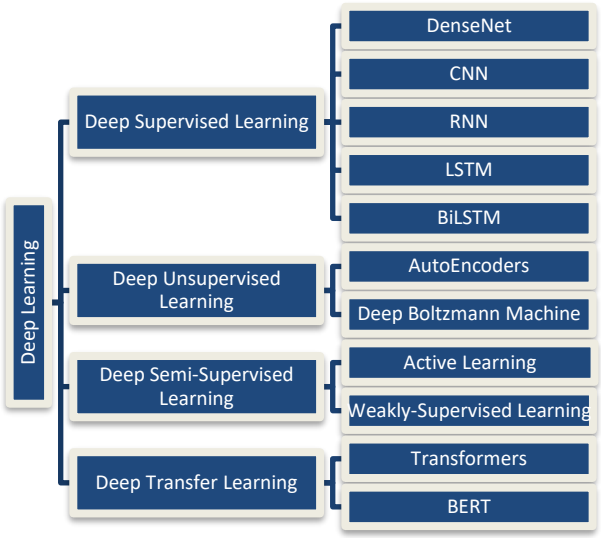


Fig. 5. Taxonomy of DL

1.2.1 Deep Supervised Learning

The DNNs for supervised learning [49] works in the same way and purpose as in the supervised learning in ML. These DNNs need the labelled data to learn the relationship and extract patterns for mapping of the input data to its corresponding output labels. These DNNs are further divided into same two types, classification, and regression same asin ML. The popular algorithms/models for DNNs for supervised learning are Dense Networks (DenseNets) [50], Convolutional Neural Networks (CNNs) [51], Recurrent Neural Networks (RNNs) [52], Long-Short Term Memory (LSTM) Networks [53], Bi-LSTM Networks [54] etc.

1.2.2 Deep Unsupervised Learning

The DNNs for unsupervised learning [55] refers to the learning where there is no labelled data. In other words, it works with the unlabelled data in the same logic as unsupervised learning in ML. The most common types of these DNNs are AutoEncoders [56] and Deep Boltzmann Machines [57].

1.2.3 Deep Semi-Supervised Learning

The DNNs have already demonstrated their performance on a large variety of deep supervised and unsupervised learning tasks (i.e. image classification [58]) once trained on extensively large labelled datasets (i.e. ImageNet [59]). However, this creates a bottleneck of creating large datasets while working with DNNs which requires extensive amount time, resources, and effort. To avoid this bottleneck, recently the DNNs for semi supervised learning [60] are introduced. The most common types of these algorithms/models are Active Learning [61] and Weakly-Supervised Learning [62].

[1] https://www.ibm.com/cloud/learn/neural-networks#toc-what-are-n-2oQ5Vepe

1.3 Natural Language Processing (NLP)

The NLP field, also referred as computational linguistics, is the subfield of AI which enables the computers to process and understand the natural language (i.e. human language) in the same way as human do. The natural language is usually in the form of free text. The goal of the NLP field is to read, decode, understand, and extract valuable, sensible insights and information from the human language. One of the scopes of this paper is to explain different methods for analyzing social media data. Therefore, it is worth to explore the field of NLP as most of the social media data is in the form of free text. The first step when it comes to analyzing unstructured text data is to convert it into the structured form such that the computers can understand that data. The common pipeline for converting unstructured text data into structured data is given in Fig. **6**.

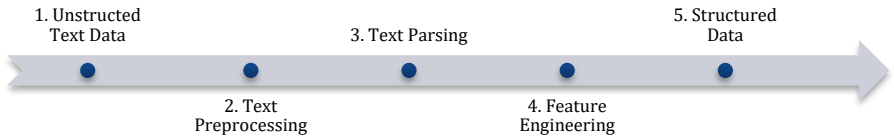


Fig. 6. From Unstructured to Structured Data using NLP

1. Unstructured Text Data

The data which does not have any defined data model or structure such that it cannot be processed easily by the computers is called unstructured text data. The most common types of such data are: online social media data, online blogs data, news websites data, electronic documents data, etc.

2. Text Preprocessing

Most of the times the unstructured text data has a lot of noise and irrelevant information which do not contribute for valuable insights and information extraction. In NLP, we have very good techniques available that can be used to pre-process/clean the unstructured text data. The different text pre-processing techniques are lowercase conversion, stemming/lemmatization, spelling correction, URLs/stopwords/puctuations removal, HTML tags removal, etc. The details for each of the techniques are explained in [63-65]. Most of the programming languages like python, JAVA, etc have built-in NLP libraries working with text pre-processing. The famous NLP libraries for python language are NLTK[i], CoreNLP[ii], Gensim[iii], Spacy[iv] and Pattern[v]. Moreover, the famous NLP libraries for JAVA are OpenNLP[vi], Stanford CoreNLP[vii] and Freeling[viii].

3. Text Parsing

Text parsing is a technique of NLP for understanding the unstructured text data. When it comes to understanding, it involves two types of techniques: Syntactic Analysis and Semantic Analysis. This section will only discuss syntactic analysis because text parsing is a syntactic analysis technique. The term "syntactic" is derived from the word "syntax". Every language has its own syntax which defines its grammatical structure while writing the text. The syntactic analysis/text parsing techniques are used to understand the grammatical structure of the human language based on formal grammar rules and meaningfulness [66]. The computer program which performs the text parsing is called "Text Parser". The most common types of text parsing for NLP are parts of speech (POS) tagging [67], shallow parsing [68], constituency parsing [69] and dependency parsing [63].

4. Feature Engineering

Feature Engineering is a technique used in NLP for understanding information from the raw text. This technique is mainly used to convert raw text into numeric form so that it can be further processed and understood by the computers while performing any task [70]. As this is a core step while converting the text into numerics so maintaining the same meaning and scope is very important here. Each numeric feature value is the representation of words and their relationships within the raw text. Fig. 7 shows an example of feature engineering for raw data. The main categories of the features are: **1) Meta features [71]:** features like no. of words in a text, no. of unique words in a text, no. of characters in a text, average length of a text, average length of words, etc. lies under this category, **2) Text-based features [72]:** The common types of text-based features are: Bag-of-

words [73], Tf-Idf [74], N-grams [75], and CountVectorizer [76] and **3) Semantic/contextual features [77]:** As compared to text-based and meta features, the semantic features help to extract this contextual meaning from the text easily. Word2Vec [77] and Doc2Vec [78] are the very first types of these features. In NLP terminology, these semantic features are often termed as "Word Embeddings" as well. The recent types of these features are FastText [79], Glove [80] and Elmo [81].

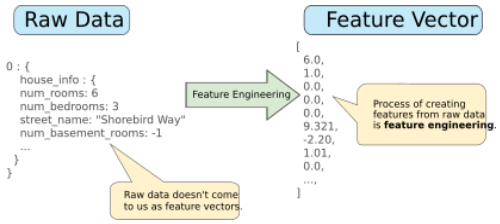


Fig. 7. Example of Feature Engineering[2]

5. Structured Data

Once, the feature engineering is completed the raw text data (i.e. unstructured data) will be in form of numbers and representing the structured data.

Next, we will discuss how the different techniques available in NLP are used for analyzing social media text to extract useful insights and information. These extracted insights help to understand user perceptions related to the domain. We will also discuss how NLP is combined with ML and DL models to extract insights from social media data. Fig. 8 shows the detailed taxonomy of some of the important NLP techniques used for perception extraction from the social media data.

Next, we will discuss different NLP techniques available in the literature for perception extraction from online social media data. Fig. 8 shows the detailed taxonomy of widely used NLP techniques for perception extraction.

1.3.1 NLP Techniques for perception extraction from social media data

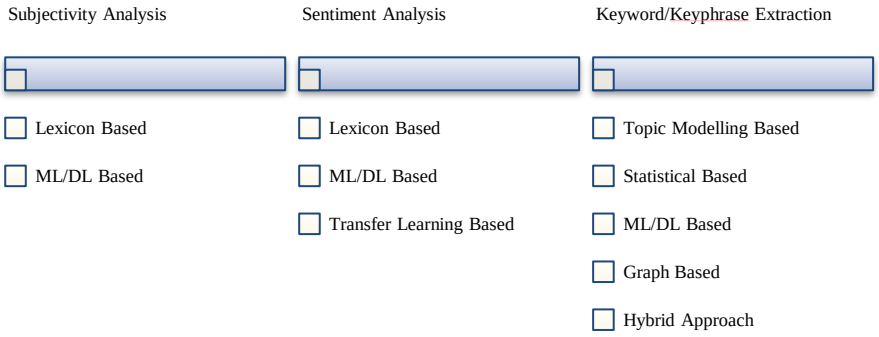


Fig. 8. Taxonomy of NLP techniques for perception extraction

I. Sentiment Analysis Techniques

Sentiment analysis (SA) also referred as Opinion Mining (OM) is a technique to extract and analyze people's opinions, attitudes, behaviours, perceptions, etc. towards different topics, products, issues being discussed on social media platforms. It is a powerful technique for businesses, industries, governments and other entities to extract and understand users mood and views [82]. The general sentiment analysis assesses the data in form of positive, negative, or neutral. However, there is more granular type of sentiment analysis used which is "emotion analysis". The emotion analysis

[2] https://towardsdatascience.com/how-to-turn-text-into-features-478b57632e99

tends to identify different emotions like anger, fear, sadness, surprise, joy, etc. from the social media text [83]. On social media platforms, people are free to express their opinions and perceptions on wide range of topics. To perform sentiment and emotion analysis on those opinions and feedbacks help to understand the users views and perceptions [84].

Next, we will discuss what are the different approaches available in the literature for performing sentiment analysis for social media data. Fig. 9 shows the taxonomy of different techniques available for performing sentiment analysis.

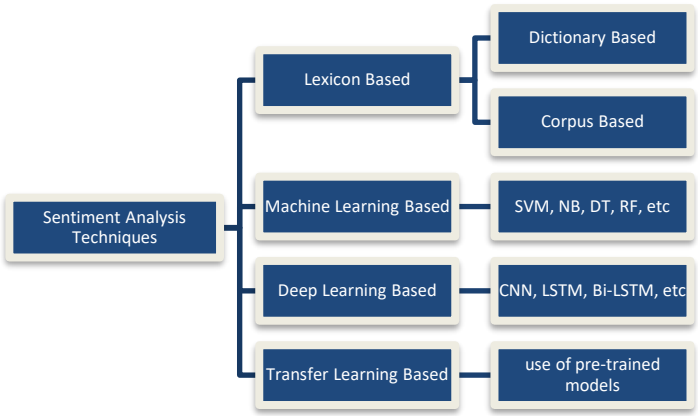


Fig. 9. Taxonomy of sentiment analysis techniques

Lexicon-based approaches are the simple dictionary-based approaches where different words and phrases are already labelled with different sentiment scores. The overall sentiment score of a text depends on the collective scores of individual words and phrases. Christopher et al. in [85] performed comparison of six different lexicon-based approaches to perform sentiment analysis. The authors evaluated all the six approaches against manually labelled amazon reviews dataset. Although, the authors achieved good accuracy in the range of 75%-77% using Hu & Liu lexicon **but the problem with this approach is that it is very limited in the scope and only applicable for the domain of product reviews.**

To overcome the lexicon-based approach challenges, several authors used machine learning based approaches for performing sentiment analysis. To perform machine learning approach, the text data needs to be converted into the numeric data according to the steps provided in the section "Natural Language Processing (NLP)". Ye et al. in [86] applied supervised machine learning algorithms namely, SVM, NB and N-gram model on yahoo reviews of famous travel destinations. The authors achieved overall 87% accuracy with N-gram model. The authors in [87] performed a systematic literature review for different machine learning models applied for performing sentiment analysis for online reviews data. **The main limitation with supervised machine learning model is the availability of already labelled training data into sentiment labels. Another limitation of using machine learning is the manual/hand crafted feature engineering for text data.s**

To solve the manual feature engineering problem recently many researchers have applied the deep learning approaches to perform sentiment analysis in social media text. The deep learning algorithms automatically identify and extract important features from the text. Pasupa et al. in [88] performed sentiment analysis using three deep learning models: CNN, LSTM and BiLSTM. The authors observed overall accuracy of 81% using CNN deep learning model. **The major limitation with using deep learning approaches is again the need of large, labelled training dataset for performing sentiment analysis in specific domains.**

II. Topic Modelling

Topic modelling is one of the very excellent technique used in NLP to understand the text. It helps to understand the text in terms analyzing and extraction of its topics. The process of learning, identifying, and extracting the topics from the text is known as "topic modelling". The

extraction of topics from the unstructured text is beneficial for several purposes such as organizing the text on the web with similar topics. Also news agencies use topic modelling to recommend articles to the readers as well. On the online social media platforms, this technique helps to extract and understand different hot trends and topics about which users are speaking through their posts, tweets, etc. Boon et al. in [89] applied topic modelling on Twitter data in order to understand what users are talking about the COVID-19 pandemic. Fig. 10 shows the word cloud generated by the authors, showing the main topics related to COVID-19 extracting using topic modelling on twitter data.

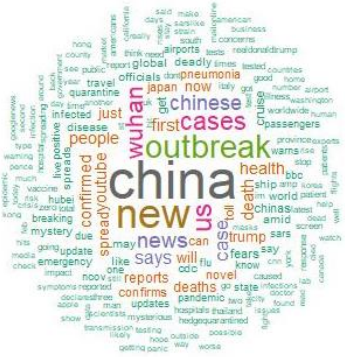


Fig. 10. Word cloud showing extracted topics from the twitter against COVID-19 [89]

There are several approaches proposed by authors in [90-92] for performing topic modelling on unstructured text data. The most widely used approaches for topic modelling from the literatures are: 1) Latent Semantic Analysis (LSA) [90], 2) Non-Negative Matrix Factorization (NNMF) [91], Probabilistic Latent Semantic Analysis (PLSA) [92] and 4) Latent Dirichlet Allocation (LDA) [93].

Although, the traditional approaches show very promising results and were used in variety of studies for performing topic modelling on social media data [94-97]. However, these approaches suffer from variety of limitations such as:

- Limit in the number of topics to be detected
- Generate generic topics
- There is trouble in dealing with topics that have complex generalization-specialization relationships.
- Less relevance of the generated topics to topics in the real-world
- Do not capture topic-inherent properties

To overcome above shortcomings, recently new techniques for topic modelling in social media data analysis are evolved based on the advancements of machine and deep learning models. These approaches are classified into two types: 1) supervised topic modelling, 2) unsupervised topic modelling [98]. The studies focusing on these techniques are discussed in the section 3.

III. Keyword/Keyphrase Extraction

Keyword/Keyphrase extraction is a NLP technique for extracting important words/concepts, relevant to the underlying document/text [99]. This technique helps in several ways like grouping of texts based on similar keywords, important concepts/topics being discussed in the texts, document searching and filtering based on extracted keywords, etc [100]. Generally, this extraction is a two-way process: 1) to generate candidate words from the text, 2) to select and rank the candidate words based on their impact on the text [101]. Fig. 11 shows the common SoTA approaches for this technique.

Fig. 11. Keyword/Keyphrase Extraction SOTA Approaches

Statistical approach: This approach generates the candidate keywords based on statistics from the text like word frequency, probability, other feature engineering techniques [102] like Tf-Idf [103], N-gram [104] and common word occurrences [105]. One of the popular algorithms using this approach is YAKE [106].

ML + DL Approach: In this approach, generally a ML/DL classifier is trained on a labelled keyword/keyphrase documents where an extracted keyword is labelled as relevant keyword or not. One of the traditional keyword extraction system based on this approach is KEA [107], which uses TF-IDF scores along with NB classifier to predict whether a candidate sentence is a keyword/keyphrase or not.

Graph-based approach: This approach generates a graph of keywords/keyphrases related to each other from the text/document. The graph connects co-occurring terms in the text with each other. The famous algorithms using this approach are TextRank [108] and RAKE [109].

Hybrid approach: This approach generates the candidate keywords/keyphrases based on combination of one or more approaches from the above. For example: ExpandRank [110], which is a combination of TF-IDF + graph-based approaches.

2 State-of-the-art studies of AI techniques for analyzing social media data

This section explains various state-of-the-art studies related to the different AI techniques explained in the section 2. Specifically, this section highlights the different models/algorithms/techniques used to analyze social media data from different social media platforms for different application areas using relevant datasets.

2.1 Machine Learning (ML)

Singh et al. in [111] analyzed the twitter data to understand the behavior of spammers distributing pornographic content on social media platform using RF machine learning algorithm. The authors reported overall accuracy of 91.96% for predicting pornographic content from Twitter data. Vafeiadis et al. in [112] conducted a comparative study of applying machine learning methods to understand customer behavior for churn prediction on a churn dataset from UCI ML repository. The authors used RF, NB, DT, LR and SVM models, out of which SVM outperformed with overall accuracy of 97%. Table. ***2*** shows additional studies where ML techniques are applied on social media data into various domains for different tasks.

Table. 2 SoTA ML techniques for social media data analysis

#	Title	ML Model Used	Dataset Used	Social Media Platform Used	Application Area
1	Crowdsourcing and collaborative learning environments based on SM [113]	Gaussian Naïve Bayes	Twitter, Facebook, LinkedIn	Twitter, Facebook, LinkedIn	Business Intelligence
2	Data analytic learning with strategic decision making [114]	DT	Twitter hashtag, Meme tracker, and Yelp	Twitter	Business Intelligence

3	Fake profile detection [115]	MRF	Facebook	Facebook	Crime detection
4	Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder [116]	Bow, SVM, LDA	Twitter, Myspace	Twitter, Myspace	Crime detection
5	Identifying Epidemics [117]	SVM, NB, and RF	Weibo		Epidemics
6	Detection of influenza epidemics [118]	Linear Regression, Multiple Regression	Twitter	Twitter	Epidemics
7	Disaster management using SM [119]	GIS model	Satellite images		Event detection
8	Real time crisis mapping of natural disasters using social media [120]	TRIDEC project	Twitter, Google Earth	Twitter	Image analysis
9	Generating person-specific representations used to boost face recognition performance [121]	SVM, LDA	PubFig83		Image analysis
10	Improving information diffusion in SM [122]	Independent cascade (IC) model and the linear threshold (LT) model	Douban, AMiner, DBLP, and LiveJournal		Recommenders' systems

2.2 Deep Learning (DL)

Untawale et. al in [123] performed age groups classification (i.e. general, teenager and adult) on twitter social media platform for tweets related to the medical domain. The authors applied MLP, DCNN, DT, RF and SVM models, out of which DCNN achieved the highest F1-score of 0.93. Guimaraes et al. in [124] applied deep learning for clustering/grouping (i.e. DeepLCRank) of the holiday photo images from Flickr and Youtube. Table. ***3*** shows additional studies where DL techniques are applied on social media data into various domains for different tasks.

Table. 3 SoTA DL techniques for social media data analysis

Title	DL Model Used	Dataset Used	Social Media Platform Used	Year	Application Area
Deep Learning for Hate Speech Detection in Tweets [125]	FastText + CNN, LSTM	Racist tweets dataset	Twitter	2017	Hate Speech
Detecting Offensive Language in Tweets Using Deep Learning [126]	RNN	Hate speech tweets	Twitter	2018	Hate Speech
Multi-layers Convolutional Neural Network for Twitter Sentiment Ordinal Scale Classification [127]	CNN	SemEval challenge dataset[3]	Twitter	2018	Sentiment Analysis
Bloom's Learning Outcomes' Automatic Classification Using LSTM and Pretrained Word Embeddings [65]	FastText + LSTM	Course learning outcomes dataset	Twitter	2021	Bloom's Taxonomy
Evaluating Polarity Trend Amidst the Coronavirus Crisis in Peoples' Attitudes toward the Vaccination Drive [128]	FastText + LSTM	Covid-19 tweets	Twitter	2021	Sentiment Analysis

[3] http://alt.qcri.org/semeval2016/task4/index.php%3fid%3ddata-and-tools

Deep learning-based personality recognition from text posts of online social networks [129]	CNN, RNN	Facebook posts	Facebook	2018	Personality Recognition
Personality recognition from Facebook text for Portuguese language [130]	LSTM	Facebook posts	Facebook	2018	Personality Recognition

2.3 Natural Language Processing (NLP)

Shamantha et al. in [131] performed sentiment analysis using ML and NLP techniques on twitter data using three ML models: NB, SVM and RF. Out of all models, RF model outperformed with overall accuracy of 80% to predict sentiments from the twitter data. Tembhurnikar et al. in [132] performed topic modelling on twitter data using sentiment analysis and n-gram approaches along with K-means algorithm to understand important topics related to events like land acquisition bills, swine fight model, etc. Table. 4shows additional studies where NLP techniques are applied by combining with different ML/DL algorithms on social media data into various domains for different tasks including sentiment analysis, topic modelling, intent detection, keyword/keyphrase extraction.

4. Findings from state-of-the-art studies

This section discusses the overall findings from the state-of-the-art studies applied for social media data analysis using ML/DL and NLP techniques for perception extraction.

Fig. 12 summarizes the findings of state-of-the-art studies for NLP techniques like sentiment analysis, topic modelling, intent detection, etc. applied in several domains like twitter, health, disaster management and news. We can further observe that for feature extraction techniques NLP has came along a long way from simple features like n-gram, Tf-Idf to more complex features like word embeddings, Elmo and Bert to understand the more complex semantics involved in the raw text. Again, for learning algorithms from past several years NLP techniques are used in combination with from simple ML algorithms like SVM, RF, NB, etc. to more complex deep learning and transformers algorithms like DNN, CNN, LSTM and BERT.

Table. 4 SoTA NLP techniques for social media data analysis

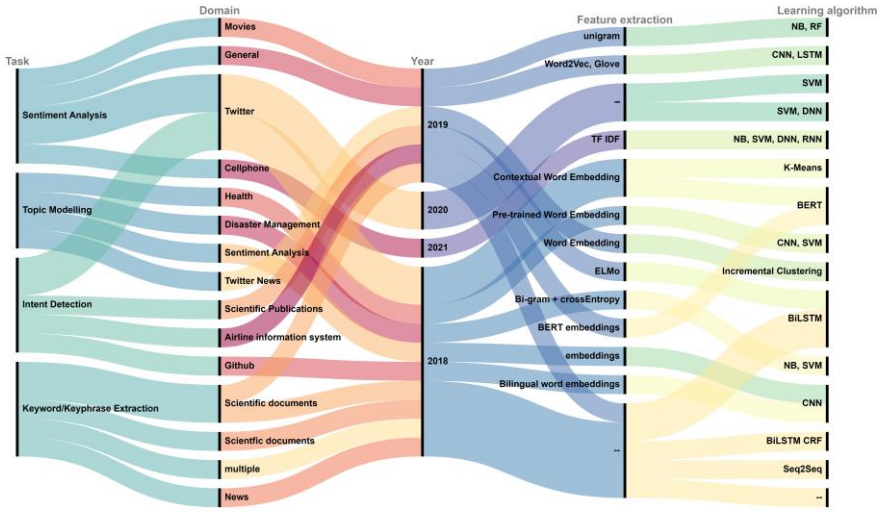


Fig. 12. Findings from SoTA NLP studies

Fig. 13 summarizes state-of-the-art studies for ML/DL algorithms applied in several domains like business analytics, epidemics, recommendation systems, crime detection, etc. From the figure, we can observe that there is a shift of applying DL algorithms as compared to ML algorithms in recent years for domains like business intelligence, hate speech on social media platforms. One possible reason might be that DL algorithms are more context-aware and semantically rich while understanding the raw text.

Reference	Type	Feature extraction	Learning Algorithm	Domain	Year	Task
Implementation of sentiment classification of movie reviews by supervised machine learning approaches [133]	Supervised	unigram	NB, RF	Movies	2019	Sentiment Analysis
Evaluation of deep learning techniques in sentiment analysis from twitter data [134]	Supervised	Word2Vec, Glove	CNN, LSTM	General	2019	Sentiment Analysis
Machine learning based aspect level sentiment analysis for Amazon products [135]	Supervised	–	SVM	Twitter	2020	Sentiment Analysis
Experimental investigation of automated system for twitter sentiment analysis to predict the public emotions using machine learning algorithms [136]	Supervised	–	SVM, DNN	Twitter	2020	Sentiment Analysis
Effect of Negation in Sentences on Sentiment Analysis and Polarity Detection [137]	Supervised	TF IDF	NB, SVM, DNN, RNN	Cellphone	2021	Sentiment Analysis
Deep representation learning for clustering of health tweets [138]	Unsupervised	Contextual Word Embedding	K-Means	Health	2018	Topic Modelling
Tweets classification with bert in the field of disaster management [139]	Supervised	Contextual Word Embedding	BERT	Disaster Management	2018	Topic Modelling
Short text classification with a convolutional neural networks based method [140]	Supervised	Pre-trained Word Embedding	CNN, SVM	Sentiment Analysis	2018	Topic Modelling
Real-time event detection from the Twitter data stream using the TwitterNews+ Framework [141]	Unsupervised	Word Embedding	Incremental Clustering	Twitter News	2019	Topic Modelling
Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents [142]	Unsupervised	--	BiLSTM CRF	Scientific documents	2019	Keyword/ Keyphrase Extraction
Exploiting topic-based adversarial neural network for cross-domain keyphrase extraction [143]	Unsupervised	--	BiLSTM	Scientific documents	2018	Keyword/ Keyphrase Extraction
Bidirectional lstm recurrent neural network for keyphrase extraction [144]	Unsupervised	--	BiLSTM	Scientific documents	2018	Keyword/ Keyphrase Extraction
Semi-supervised learning for neural keyphrase generation [145]	Semi-supervised	--	Seq2Seq	multiple	2018	Keyword/ Keyphrase Extraction
Learning feature representations for keyphrase extraction [146]	Graph based	--	--	News	2018	Keyword/ Keyphrase Extraction

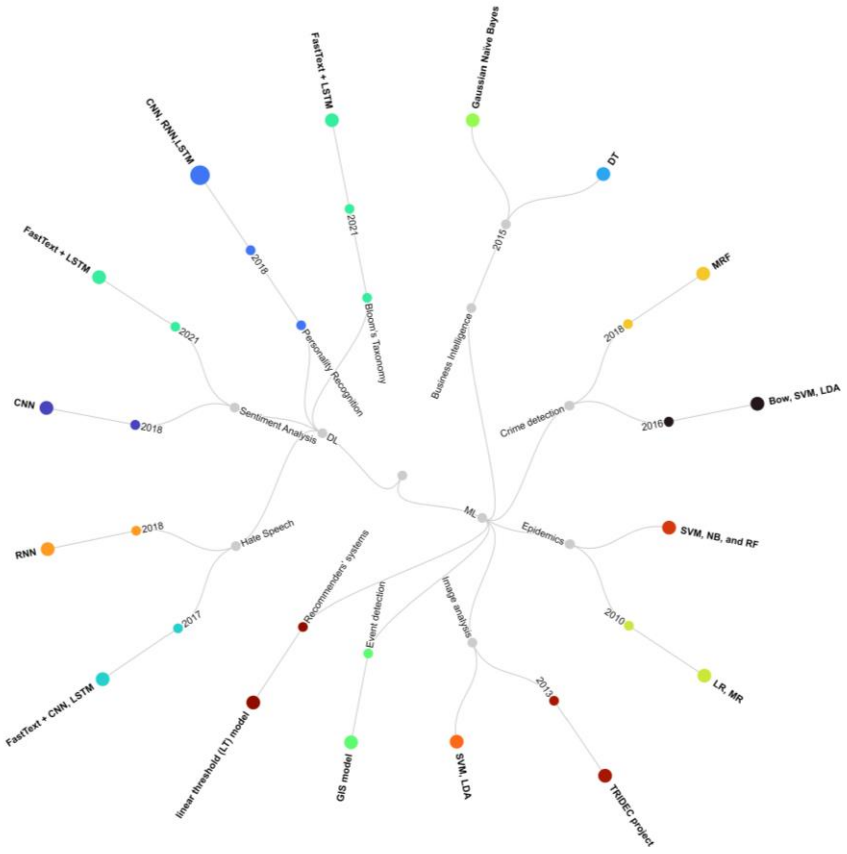


Fig. 13. Findings from SoTA ML/DL studies

5. Conclusion

In this paper, we described several artificial intelligence techniques including machine learning, deep learning and natural language processing in detail for the purpose of social media data analysis. Along with describing the techniques, we also conducted state-of-the-art review of studies where these techniques are applied for social media data analysis. The findings of review resulted in the identification of existing domains where the techniques are most widely used. Also, the review revealed the shift of the learning algorithms from ML to DL in recent studies in the year 2021. Overall, the review provides good discussions related to the different type of algorithms applied in various domains for achieving various types of tasks.

References

1. McGregor, S.C., *Social media as public opinion: How journalists use social media to represent public opinion.* Journalism, 2019. **20**(8): p. 1070-1086.
2. Sofalvi, A.J. and C.O. Airhihenbuwa, *An analysis of the relationship between news coverage of health topics and public opinion of the most important health problems in the United States.* Journal of Health Education, 1992. **23**(5): p. 296-300.
3. Krugman, H.E., *The impact of television advertising: Learning without involvement.* Public opinion quarterly, 1965. **29**(3): p. 349-356.
4. Tobin, C., A.R. Moodie, and C. Livingstone, *A review of public opinion towards alcohol controls in Australia.* BMC Public Health, 2011. **11**(1): p. 1-9.
5. D'Andrea, E., et al., *Monitoring the public opinion about the vaccination topic from tweets analysis.* Expert Systems with Applications, 2019. **116**: p. 209-226.

6. Murphy, J., et al., *Social media in public opinion research: Executive summary of the AAPOR task force on emerging technologies in public opinion research.* Public Opinion Quarterly, 2014. **78**(4): p. 788-794.
7. Fishkin, J.S., *Beyond polling alone: the quest for an informed public.* Critical Review, 2006. **18**(1-3): p. 157-165.
8. Bian, J., et al., *Mining Twitter to assess the public perception of the "Internet of Things".* PloS one, 2016. **11**(7): p. e0158450.
9. Khan, S., et al., *Antecedents of trust in using social media for E-government services: An empirical study in Pakistan.* Technology in Society, 2021. **64**: p. 101400.
10. Huang, M.-h., T. Whang, and L. Xuchuan, *The Internet, social capital, and civic engagement in Asia.* Social Indicators Research, 2017. **132**(2): p. 559.
11. Kang, Y., et al., *The public's opinions on a new school meals policy for childhood obesity prevention in the US: A social media analytics approach.* International journal of medical informatics, 2017. **103**: p. 83-88.
12. Anstead, N. and B. O'Loughlin, *Social media analysis and public opinion: The 2010 UK general election.* Journal of computer-mediated communication, 2015. **20**(2): p. 204-220.
13. Chen, S.-C. and C.-P. Lin, *Understanding the effect of social media marketing activities: The mediation of social identification, perceived value, and satisfaction.* Technological Forecasting and Social Change, 2019. **140**: p. 22-32.
14. Mollema, L., et al., *Disease detection or public opinion reflection? Content analysis of tweets, other social media, and online newspapers during the measles outbreak in The Netherlands in 2013.* Journal of medical Internet research, 2015. **17**(5): p. e3863.
15. Xue, Y., et al., *Relationship discovery in public opinion and actual behavior for social media stock data space.* EURASIP Journal on Wireless Communications and Networking, 2016. **2016**(1): p. 1-13.
16. Pourebrahim, N., et al., *Understanding communication dynamics on Twitter during natural disasters: A case study of Hurricane Sandy.* International journal of disaster risk reduction, 2019. **37**: p. 101176.
17. Adams-Cohen, N.J., *Policy change and public opinion: Measuring shifting political sentiment with social media data.* American Politics Research, 2020. **48**(5): p. 612-621.
18. Salleh, S.M., *From survey to social media: Public opinion and politics in the age of big data.* Advanced Science Letters, 2017. **23**(11): p. 10696-10700.
19. Dong, X. and Y. Lian, *A review of social media-based public opinion analyses: Challenges and recommendations.* Technology in Society, 2021. **67**: p. 101724.
20. Lokeshkumar, R., O.A. Mishra, and S. Kalra, *Social media data analysis to predict mental state of users using machine learning techniques.* Journal of Education and Health Promotion, 2021. **10**(1): p. 301.
21. Guo, Y., et al., *The application of artificial intelligence and data integration in COVID-19 studies: a scoping review.* Journal of the American Medical Informatics Association, 2021.
22. Luo, J., et al., *Exploring temporal suicidal behavior patterns on social media: Insight from Twitter analytics.* Health informatics journal, 2020. **26**(2): p. 738-752.
23. Valdez, D., et al., *Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of twitter data.* Journal of medical Internet research, 2020. **22**(12): p. e21418.
24. Santander, P., et al., *Analyzing social media, analyzing the social? A methodological discussion about the demoscopic and predictive potential of social media.* Quality & Quantity, 2020: p. 1-21.
25. Martinez, L.S., M.-H. Tsou, and B.H. Spitzberg. *A case study in belief surveillance, sentiment analysis, and identification of informational targets for e-cigarettes interventions*. in *Proceedings of the 10th International Conference on Social Media and Society*. 2019.
26. Saravanan, M. and S.K. Perepu, *Realizing social-media-based analytics for smart agriculture.* The Review of Socionetwork Strategies, 2019. **13**(1): p. 33-53.
27. Chang, Y.-C., C.-H. Ku, and C.-H. Chen, *Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor.* International Journal of Information Management, 2019. **48**: p. 263-279.
28. von Hoffen, M., et al., *Leveraging social media to gain insights into service delivery: a study on Airbnb.* Information Systems and e-Business Management, 2018. **16**(2): p. 247-269.
29. Chumwatana, T. and K. Wongkolkitsilp. *Using Classification Technique for Customer Relationship Management based on Thai Social Media Data*. in *Proceedings of the 2019 11th International Conference on Computer and Automation Engineering*. 2019.
30. Tian, X., et al., *A new approach of social media analytics to predict service quality: evidence from the airline industry.* Journal of Enterprise Information Management, 2019.
31. Dahal, B., S.A. Kumar, and Z. Li, *Topic modeling and sentiment analysis of global climate change tweets.* Social Network Analysis and Mining, 2019. **9**(1): p. 1-20.
32. Dias, D.S., M.D. Welikala, and N.G. Dias. *Identifying racist social media comments in Sinhala language using text analytics models with machine learning*. in *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*. 2018. IEEE.

33. Barrelet, C.J., S.S. Kuzulugil, and A.B. Bener. *The Twitter Bullishness Index: A Social Media Analytics Indicator for the Stock Market*. in *Proceedings of the 20th International Database Engineering & Applications Symposium*. 2016.
34. Park, S.B., J. Jang, and C.M. Ok, *Analyzing Twitter to explore perceptions of Asian restaurants.* Journal of Hospitality and Tourism Technology, 2016.
35. Pushpam, C.A. and J.G. Jayanthi, *Overview on data mining in social media.* International Journal of Computer Sciences and Engineering, 2017. **5**(11): p. 147-157.
36. Camacho, D., M.V. Luzón, and E. Cambria, *New trends and applications in social media analytics*. 2021, Elsevier.
37. Balan, S. and J. Rege, *Mining for social media: Usage patterns of small businesses.* Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy, 2017. **8**(1): p. 43-50.
38. Quinlan, J.R., *C4. 5: programs for machine learning*. 2014: Elsevier.
39. Whelan, E., A.N. Islam, and S. Brooks, *Applying the SOBC paradigm to explain how social media overload affects academic performance.* Computers & Education, 2020. **143**: p. 103692.
40. Alpaydin, E., *Introduction to machine learning*. 2020: MIT press.
41. Rebala, G., A. Ravi, and S. Churiwala, *Learning Models*, in *An Introduction to Machine Learning*. 2019, Springer International Publishing: Cham. p. 19-23.
42. Rebala, G., A. Ravi, and S. Churiwala, *Classification*, in *An Introduction to Machine Learning*. 2019, Springer International Publishing: Cham. p. 57-66.
43. Rebala, G., A. Ravi, and S. Churiwala, *Regressions*, in *An Introduction to Machine Learning*. 2019, Springer International Publishing: Cham. p. 25-40.
44. Rebala, G., A. Ravi, and S. Churiwala, *Clustering*, in *An Introduction to Machine Learning*. 2019, Springer International Publishing: Cham. p. 67-76.
45. Zhu, X. and A.B. Goldberg, *Introduction to semi-supervised learning.* Synthesis lectures on artificial intelligence and machine learning, 2009. **3**(1): p. 1-130.
46. T.K, B., C.S.R. Annavarapu, and A. Bablani, *Machine learning algorithms for social media analysis: A survey.* Computer Science Review, 2021. **40**: p. 100395.
47. Goodfellow, I., Y. Bengio, and A. Courville, *Deep learning*. 2016: MIT press.
48. Yegnanarayana, B., *Artificial neural networks*. 2009: PHI Learning Pvt. Ltd.
49. Makantasis, K., et al. *Deep supervised learning for hyperspectral data classification through convolutional neural networks*. in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. 2015. IEEE.
50. Mele, A., *A structural model of dense network formation.* Econometrica, 2017. **85**(3): p. 825-850.
51. Albawi, S., T.A. Mohammed, and S. Al-Zawi. *Understanding of a convolutional neural network*. in *2017 International Conference on Engineering and Technology (ICET)*. 2017. Ieee.
52. Merrill, W., et al., *A formal hierarchy of RNN architectures.* arXiv preprint arXiv:2004.08500, 2020.
53. Sundermeyer, M., R. Schlüter, and H. Ney. *LSTM neural networks for language modeling*. in *Thirteenth annual conference of the international speech communication association*. 2012.
54. Xu, G., et al., *Sentiment analysis of comment texts based on BiLSTM.* Ieee Access, 2019. **7**: p. 51522-51532.
55. Valpola, H., *From neural PCA to deep unsupervised learning*, in *Advances in independent component analysis and learning machines*. 2015, Elsevier. p. 143-171.
56. Lange, S. and M. Riedmiller. *Deep auto-encoder neural networks in reinforcement learning*. in *The 2010 International Joint Conference on Neural Networks (IJCNN)*. 2010. IEEE.
57. Salakhutdinov, R. and G. Hinton. *Deep boltzmann machines*. in *Artificial intelligence and statistics*. 2009. PMLR.
58. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems, 2012. **25**: p. 1097-1105.
59. Deng, J., et al. *Imagenet: A large-scale hierarchical image database*. in *2009 IEEE conference on computer vision and pattern recognition*. 2009. Ieee.
60. Ouali, Y., C. Hudelot, and M. Tami, *An overview of deep semi-supervised learning.* arXiv preprint arXiv:2006.05278, 2020.
61. Settles, B., *Active learning literature survey.* 2009.
62. Ratner, A., et al., *Weak supervision: the new programming paradigm for machine learning.* Hazy Research. Available via https://dawn. cs. stanford. edu//2017/07/16/weak-supervision/. Accessed, 2019: p. 05-09.
63. Shaikh, S. and S.M. Doudpotta, *Aspects based opinion mining for teacher and course evaluation.* Sukkur IBA Journal of Computing and Mathematical Sciences, 2019. **3**(1): p. 34-43.
64. Shaikh, S., et al., *Towards Improved Classification Accuracy on Highly Imbalanced Text Dataset Using Deep Neural Language Models.* Applied Sciences, 2021. **11**(2): p. 869.
65. Shaikh, S., S.M. Daudpotta, and A.S. Imran, *Bloom's Learning Outcomes' Automatic Classification Using LSTM and Pretrained Word Embeddings.* IEEE Access, 2021. **9**: p. 117887-117909.

66. Granaas, M.M., *Simple, applied text parsing.* Behavior Research Methods, Instruments, & Computers, 1985. **17**(2): p. 209-216.
67. Kumawat, D. and V. Jain, *POS tagging approaches: a comparison.* International Journal of Computer Applications, 2015. **118**(6).
68. Munoz, M., et al., *A learning approach to shallow parsing.* arXiv preprint cs/0008022, 2000.
69. Rodrıguez, C.G., *Parsing Schemata for Practical Text Analysis*. 2011, Citeseer.
70. Guyon, I. and A. Elisseeff, *An introduction to feature extraction*, in *Feature extraction*. 2006, Springer. p. 1-25.
71. Chen, W., et al., *Exploiting meta features for dependency parsing and part-of-speech tagging.* Artificial Intelligence, 2016. **230**: p. 173-191.
72. Blackstock, A. and M. Spitz, *Classifying movie scripts by genre with a MEMM using NLP-Based features.* Citeseer, 2008.
73. Zhang, Y., R. Jin, and Z.-H. Zhou, *Understanding bag-of-words model: a statistical framework.* International Journal of Machine Learning and Cybernetics, 2010. **1**(1-4): p. 43-52.
74. Shi, C.-y., C.-j. Xu, and X.-J. Yang, *Study of TFIDF algorithm.* Journal of Computer Applications, 2009. **29**(6): p. 167-170.
75. Pizarro, J. *Using N-grams to detect Bots on Twitter*. in *CLEF (Working Notes)*. 2019.
76. Kulkarni, A. and A. Shivananda, *Converting text to features*, in *Natural language processing recipes*. 2021, Springer. p. 63-106.
77. Church, K.W., *Word2Vec.* Natural Language Engineering, 2017. **23**(1): p. 155-162.
78. Lau, J.H. and T. Baldwin, *An empirical evaluation of doc2vec with practical insights into document embedding generation.* arXiv preprint arXiv:1607.05368, 2016.
79. Joulin, A., et al., *Fasttext. zip: Compressing text classification models.* arXiv preprint arXiv:1612.03651, 2016.
80. Pennington, J., R. Socher, and C.D. Manning. *Glove: Global vectors for word representation*. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
81. Zhu, H., I.C. Paschalidis, and A. Tahmasebi, *Clinical concept extraction with contextual word embedding.* arXiv preprint arXiv:1810.10566, 2018.
82. Birjali, M., M. Kasri, and A. Beni-Hssane, *A comprehensive survey on sentiment analysis: Approaches, challenges and trends.* Knowledge-Based Systems, 2021: p. 107134.
83. Nandwani, P. and R. Verma, *A review on sentiment analysis and emotion detection from text.* Social Network Analysis and Mining, 2021. **11**(1): p. 1-19.
84. Ahmad, Z., et al., *Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding.* Expert Systems with Applications, 2020. **139**: p. 112851.
85. Khoo, C.S. and S.B. Johnkhan, *Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons.* Journal of Information Science, 2018. **44**(4): p. 491-511.
86. Ye, Q., Z. Zhang, and R. Law, *Sentiment classification of online reviews to travel destinations by supervised machine learning approaches.* Expert systems with applications, 2009. **36**(3): p. 6527-6535.
87. Jain, P.K., R. Pamula, and G. Srivastava, *A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews.* Computer Science Review, 2021. **41**: p. 100413.
88. Pasupa, K. and T.S.N. Ayutthaya, *Thai sentiment analysis with deep learning techniques: A comparative study based on word embedding, POS-tag, and sentic features.* Sustainable Cities and Society, 2019. **50**: p. 101615.
89. Boon-Itt, S. and Y. Skunkan, *Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study.* JMIR Public Health and Surveillance, 2020. **6**(4): p. e21978.
90. Deerwester, S., et al., *Indexing by latent semantic analysis.* Journal of the American society for information science, 1990. **41**(6): p. 391-407.
91. Paatero, P. and U. Tapper, *Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values.* Environmetrics, 1994. **5**(2): p. 111-126.
92. Cooper, G.F. and S. Moral, *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. 1998: Morgan Kaufmann Publishers Inc.
93. De Finetti, B., *Theory of probability: A critical introductory treatment*. Vol. 6. 2017: John Wiley & Sons.
94. Nguyen, D.Q., et al., *Improving topic models with latent feature word representations.* Transactions of the Association for Computational Linguistics, 2015. **3**: p. 299-313.
95. Yin, H., et al. *A unified model for stable and temporal topic detection from social media data*. in *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. 2013. IEEE.
96. Xie, W., et al., *Topicsketch: Real-time bursty topic detection from twitter.* IEEE Transactions on Knowledge and Data Engineering, 2016. **28**(8): p. 2216-2229.

97. Cataldi, M., L. Di Caro, and C. Schifanella. *Emerging topic detection on twitter based on temporal and social terms evaluation*. in *Proceedings of the tenth international workshop on multimedia data mining*. 2010.
98. Mottaghinia, Z., et al., *A review of approaches for topic detection in Twitter.* Journal of Experimental & Theoretical Artificial Intelligence, 2021. **33**(5): p. 747-773.
99. Siddiqi, S. and A. Sharan, *Keyword and keyphrase extraction techniques: a literature review.* International Journal of Computer Applications, 2015. **109**(2).
100. Lahiri, S., S.R. Choudhury, and C. Caragea, *Keyword and keyphrase extraction using centrality measures on collocation networks.* arXiv preprint arXiv:1401.6571, 2014.
101. Zhao, D., et al. *Keyword extraction for social media short text*. in *2017 14th Web Information Systems and Applications Conference (WISA)*. 2017. IEEE.
102. Luhn, H.P., *A statistical approach to mechanized encoding and searching of literary information.* IBM Journal of research and development, 1957. **1**(4): p. 309-317.
103. Salton, G. and C. Buckley, *Term-weighting approaches in automatic text retrieval.* Information processing & management, 1988. **24**(5): p. 513-523.
104. Cohen, J.D., *Highlights: Language-and domain-independent automatic indexing terms for abstracting.* Journal of the American society for information science, 1995. **46**(3): p. 162-174.
105. Matsuo, Y. and M. Ishizuka, *Keyword extraction from a single document using word co-occurrence statistical information.* International Journal on Artificial Intelligence Tools, 2004. **13**(01): p. 157-169.
106. Campos, R., et al. *Yake! collection-independent automatic keyword extractor*. in *European Conference on Information Retrieval*. 2018. Springer.
107. Witten, I.H., et al., *Kea: Practical automated keyphrase extraction*, in *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*. 2005, IGI global. p. 129-152.
108. Mihalcea, R. and P. Tarau. *Textrank: Bringing order into text*. in *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004.
109. Mothe, J., F. Ramiandrisoa, and M. Rasolomanana. *Automatic keyphrase extraction using graph-based methods*. in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. 2018.
110. Wan, X. and J. Xiao. *Single Document Keyphrase Extraction Using Neighborhood Knowledge*. in *AAAI*. 2008.
111. Singh, M., D. Bansal, and S. Sofat, *Behavioral analysis and classification of spammers distributing pornographic content in social media.* Social Network Analysis and Mining, 2016. **6**(1): p. 1-18.
112. Vafeiadis, T., et al., *A comparison of machine learning techniques for customer churn prediction.* Simulation Modelling Practice and Theory, 2015. **55**: p. 1-9.
113. Stantchev, V., L. Prieto-González, and G. Tamm, *Cloud computing service for knowledge assessment and studies recommendation in crowdsourcing and collaborative learning environments based on social network analysis.* Computers in Human Behavior, 2015. **51**: p. 762-770.
114. Chen, Y., et al., *Decision learning: Data analytic learning with strategic decision making.* IEEE Signal Processing Magazine, 2015. **33**(1): p. 37-56.
115. Ramalingam, D. and V. Chinnaiah, *Fake profile detection techniques in large-scale online social networks: A comprehensive review.* Computers & Electrical Engineering, 2018. **65**: p. 165-177.
116. Zhao, R. and K. Mao, *Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder.* IEEE Transactions on Affective Computing, 2016. **8**(3): p. 328-339.
117. Li, L., et al., *Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo.* IEEE Transactions on Computational Social Systems, 2020. **7**(2): p. 556-562.
118. Culotta, A. *Towards detecting influenza epidemics by analyzing Twitter messages*. in *Proceedings of the first workshop on social media analytics*. 2010.
119. Xu, Z.-X., Y. Liu, and J. Zhang, *A pair of novel 4-connected homochiral coordination polymers based on proline-tetrazole ligand.* Inorganic Chemistry Communications, 2016. **67**: p. 44-46.
120. Middleton, S.E., L. Middleton, and S. Modafferi, *Real-time crisis mapping of natural disasters using social media.* IEEE Intelligent Systems, 2013. **29**(2): p. 9-17.
121. Chiachia, G., et al., *Learning person-specific representations from faces in the wild.* IEEE Transactions on Information Forensics and Security, 2014. **9**(12): p. 2089-2099.
122. Wang, Z., et al., *Activity maximization by effective information diffusion in social networks.* IEEE Transactions on Knowledge and Data Engineering, 2017. **29**(11): p. 2374-2387.
123. Guimaraes, R.G., et al., *Age groups classification in social network using deep learning.* IEEE Access, 2017. **5**: p. 10805-10816.
124. Zin, T.T., P. Tin, and H. Hama. *Deep learning model for integration of clustering with ranking in social networks*. in *International Conference on Genetic and Evolutionary Computing*. 2016. Springer.
125. Badjatiya, P., et al. *Deep learning for hate speech detection in tweets*. in *Proceedings of the 26th international conference on World Wide Web companion*. 2017.

126. Pitsilis, G.K., H. Ramampiaro, and H. Langseth, *Detecting offensive language in tweets using deep learning.* arXiv preprint arXiv:1801.04433, 2018.
127. Alali, M., et al. *Multi-layers convolutional neural network for twitter sentiment ordinal scale classification*. in *International Conference on Soft Computing and Data Mining*. 2018. Springer.
128. Batra, R., et al., *Evaluating Polarity Trend Amidst the Coronavirus Crisis in Peoples' Attitudes toward the Vaccination Drive.* Sustainability, 2021. **13**(10): p. 5344.
129. Xue, D., et al., *Deep learning-based personality recognition from text posts of online social networks.* Applied Intelligence, 2018. **48**(11): p. 4232-4246.
130. da Silva, B.B.C. and I. Paraboni. *Personality recognition from Facebook text*. in *International Conference on Computational Processing of the Portuguese Language*. 2018. Springer.
131. Shamantha, R.B., S.M. Shetty, and P. Rai. *Sentiment analysis using machine learning classifiers: evaluation of performance*. in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*. 2019. IEEE.
132. Tembhurnikar, S.D. and N.N. Patil. *Topic detection using BNgram method and sentiment analysis on twitter dataset*. in *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions)*. 2015. IEEE.
133. Untawale, T.M. and G. Choudhari. *Implementation of sentiment classification of movie reviews by supervised machine learning approaches*. in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. 2019. IEEE.
134. Goularas, D. and S. Kamis. *Evaluation of deep learning techniques in sentiment analysis from twitter data*. in *2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML)*. 2019. IEEE.
135. Nandal, N., R. Tanwar, and J. Pruthi, *Machine learning based aspect level sentiment analysis for Amazon products.* Spatial Information Research, 2020. **28**(5): p. 601-607.
136. Sharma, P. and A. Sharma, *Experimental investigation of automated system for twitter sentiment analysis to predict the public emotions using machine learning algorithms.* Materials Today: Proceedings, 2020.
137. Mukherjee, P., et al., *Effect of Negation in Sentences on Sentiment Analysis and Polarity Detection.* Procedia Computer Science, 2021. **185**: p. 370-379.
138. Gencoglu, O., *Deep representation learning for clustering of health tweets.* arXiv preprint arXiv:1901.00439, 2018.
139. Ma, G., *Tweets classification with bert in the field of disaster management.* Dept. Civil Eng., Stanford Univ., Stanford, CA, USA, Tech. Rep, 2019. **15785631**.
140. Hu, Y., et al. *Short text classification with a convolutional neural networks based method*. in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. 2018. IEEE.
141. Hasan, M., M.A. Orgun, and R. Schwitter, *Real-time event detection from the Twitter data stream using the TwitterNews+ Framework.* Information Processing & Management, 2019. **56**(3): p. 1146-1165.
142. Alzaidy, R., C. Caragea, and C.L. Giles. *Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents*. in *The world wide web conference*. 2019.
143. Wang, Y., et al. *Exploiting topic-based adversarial neural network for cross-domain keyphrase extraction*. in *2018 IEEE International Conference on Data Mining (ICDM)*. 2018. IEEE.
144. Basaldella, M., et al. *Bidirectional lstm recurrent neural network for keyphrase extraction*. in *Italian Research Conference on Digital Libraries*. 2018. Springer.
145. Ye, H. and L. Wang, *Semi-supervised learning for neural keyphrase generation.* arXiv preprint arXiv:1808.06773, 2018.
146. Florescu, C. and W. Jin. *Learning feature representations for keyphrase extraction*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018.

[i] https://www.nltk.org/
[ii] https://github.com/stanfordnlp/python-stanford-corenlp
[iii] https://pypi.org/project/gensim/
[iv] https://spacy.io/
[v] https://analyticsindiamag.com/hands-on-guide-to-pattern-a-python-tool-for-effective-text-processing-and-data-mining/
[vi] https://opennlp.apache.org/
[vii] https://stanfordnlp.github.io/CoreNLP/
[viii] http://nlp.lsi.upc.edu/freeling/node/1