Survey of Text Mining Algorithms and Techniques for Social Media Analysis

SANIDHYA VIJAYVARGIYA, 2020A7PS2056H, BITS Pilani Hyderabad Campus, India
MEDINI N B, 2020A7PS1722H, BITS Pilani Hyderabad Campus, India
MEGHA KHURANA, 2020A7PS1316H, BITS Pilani Hyderabad Campus, India
NISHITH KUMAR, 2020A7PS0157H, BITS Pilani Hyderabad Campus, India

The COVID-19 pandemic has brought about a revolutionary change to the daily lives of individuals. It has forced society to virtualize everyday tasks like schooling, shopping, and working at the office. The "new normal" has left many disoriented, and people are still struggling to adapt to this reality. The impact of these drastic changes to daily lives combined with a lack of company to express one's thoughts and emotions due to lockdown has led to an unprecedented influx of new social media users. These new users consist of demographics such as middle-aged and senior citizens who have had to adapt to social media as the primary means of exchanging information. Access to the internet has become more affordable as well. Social media apps like Twitter have become the primary platform for political discourse and social activism. This has led to the data generated from social media capturing a more accurate representation of society. Analysis of this data holds the potential to provide more accurate insights due to its vastness. Because of their widespread use, social networking services generate vast amounts of data, characterized by three computing issues: volume, noise, and dynamism. Because of these challenges, manually analyzing social network data is typically difficult, necessitating the use of data mining methods. Text mining is the extraction of information from unstructured text by converting it to normalized, structured data. Machine learning techniques are used in conjunction with text mining techniques to derive insightful analyses from the data. Text mining plays a crucial role when it comes to analyzing data from social media platforms as spelling, grammar, and sentence structure are frequently overlooked in everyday conversations. This creates problems in lexical, syntactic, and semantic analysis, making it challenging to find discernable patterns in the data. To handle the vast amount of data, the text mining algorithms need to be chosen wisely based on the requirements– balancing the training time and the memory required. There are five commonly used and efficient techniques in text mining- Information Extraction, Information Retrieval, Categorization, Clustering, and Summarization. Text mining often requires to be used in conjunction with Natural Language Processing technologies to parse and interpret datasets. With the rise of less supervised techniques, text mining software is now better able to find hidden similarities and relationships in text data. The goal of this paper is to analyze text mining techniques focusing on their effectiveness for social media applications. We take a detailed look at various challenges faced in text mining and how the existing algorithms aim to overcome them. We hope to provide direction to future researchers by presenting them with the advantages and limitations of various text mining algorithms and practices. More research is required on the COVID-19 pandemic data to exhaustively compare the performance of the text mining algorithms with their performance on data before COVID-19.

Additional Key Words and Phrases: Social media, Text mining, COVID-19, Data Mining

Authors' addresses: Sanidhya Vijayvargiya, 2020A7PS2056H, BITS Pilani Hyderabad Campus, Hyderabad, India, f20202056@hyderabad.bits-pilani.ac.in; Medini N B, f20201722@hyderabad.bits-pilani.ac.in, 2020A7PS1722H, BITS Pilani Hyderabad Campus, Hyderabad, India; Megha Khurana, f20201316@hyderabad.bits-pilani.ac.in, 2020A7PS1316H, BITS Pilani Hyderabad Campus, Hyderabad, India; Nishith Kumar, f20200157@hyderabad.bits-pilani.ac.in, 2020A7PS0157H, BITS Pilani Hyderabad Campus, Hyderabad, India.

Manuscript submitted to ACM

ACM Reference Format:
Sanidhya Vijayvargiya, Medini N B, Megha Khurana, and Nishith Kumar. 2022. Survey of Text Mining Algorithms and Techniques for Social Media Analysis. 1, 1 (September 2022), 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

A social network is a web-based service that allows users to create a public/semi-public profile and share information and interact with other users within the network. Social networks are important sources of online interactions and contents sharing, subjectivity, assessments, approaches, evaluation, influences, observations, feelings, opinions, and sentiments expressions borne out in text, discussions, reviews, blogs, remarks, news, reactions, or some other documents. 1 Social network analysis is the study of social networks to understand their structure and the behavior of their users in

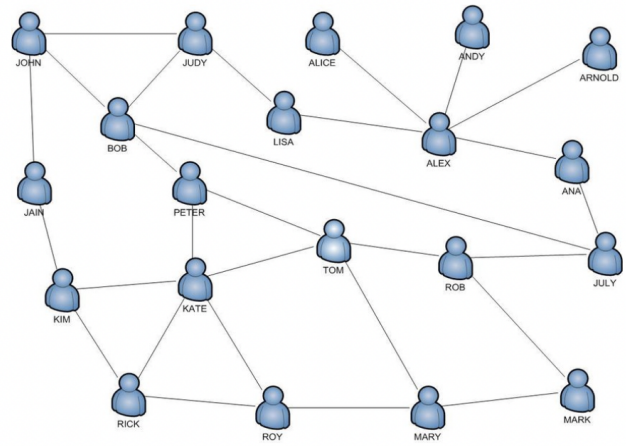

Fig. 1. Social media network

an online setting. It has proved its importance by its use in different fields- like product management, search engines, social analytics, expert finding, image analysis, fraud detection, viral marketing, etc.

A social network can be represented by a graph, where the nodes represent entities and the edges represent links between them. These networks can also be used to represent many real-world aspects like phone calls and power grids, apart from their social aspect.

The different types of social media analysis are[11]:

- Socio-Centric Network Analysis:
  It emerged in sociology, anthropology, and psychology.
  It involves the analysis of interactions between an individual(called ego) and other people related to the ego.
- Knowledge-Based Network Analysis:
  It emerged in computer science.
  It involves the quantification of interaction between individuals, groups, and other entities.

  Data Mining is the process of uncovering patterns and other hidden information from large datasets. There has been a recent surge of interest, amongst the data mining community, to uncover important hidden patterns

from the large amount of data gained from social networks. Data Mining can be used to build descriptive and predictive models of social interactions.

However, traditional data mining techniques fail when applied to social media, due to its highly dynamic and noisy nature. Thus, different and unconventional methods need to be designed for social media, which we will be discussing in this survey paper. We'll be discussing various methods of topic detection, followed by methods of opinion analysis from social media microblogging and sentiment analysis.

2 SENTIMENT ANALYSIS ON SOCIAL NETWORK

Sentiment analysis is a type of text mining that discovers and extracts subjective information from a source document. It focuses on a text's polarity (positive, negative, or neutral), but it can also detect specific emotional states, urgency, and even intentions beyond polarity. Analyzing the sentiment of messages shared on social media can have a plethora of uses[51]. Firstly, it can help analyze the reaction of the public to an event or an action by the government. Secondly, increasing barbaric use of social media by politically motivated individuals or otherwise and the lack of an effective surveillance system have motivated research on how certain users attempt to force their views on others and spread propaganda. Such tweets are called highly polar and should not be used as a primary source of information for the general public. Other applications include measuring brand health, identifying emerging trends, getting feedback, and resolving problems faced by customers[35].

The goal of sentiment analysis on social networks is to identify a possible shift in society in terms of stakeholder or public opinions, observations, and expectations. This recognition allows the organizations in question to act quickly by making appropriate decisions. It is critical to convert sentiment into valuable knowledge through mining and analysis. Liu[52] summarized all relevant research subjects in the field of sentiment analysis, involving more than 400 bibliographic references from prominent journals and conferences, out of all the surveys proposed in recent years. There are two primary methods for sentiment analysis[13], namely, lexicon-based and machine learning-based. 2

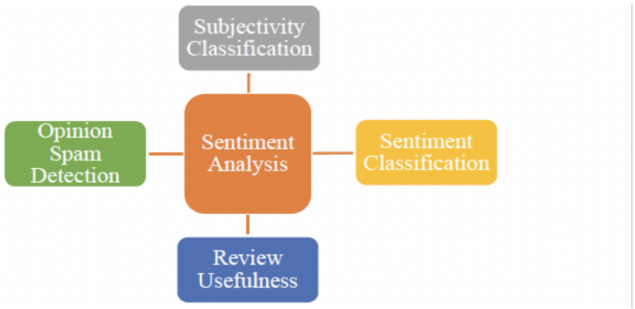

Fig. 2. Sentiment analysis tasks

2.1 Lexicon-based sentiment analysis

The lexicon-based method is one of the approaches or techniques used in semantic analysis. Unsupervised sentiment analysis majorly relies on this mode of sentiment analysis. From the sentiment polarity of lexicons, this technique estimates the sentiment polarities of the entire document or set of sentences. Positive, negative, or neutral semantic orientation is possible. Both manually and programmatically developed lexical dictionaries are available. Many researchers

use the WorldNet dictionary. To begin, lexicons are extracted from the entire document, and then WorldNet or another online thesaurus can be utilized to find synonyms and antonyms to expand the vocabulary.

Adjectives and adverbs are used in lexicon-based procedures to determine the text's sentiment polarity. Adjective and adverb combinations are retrieved with their sentiment orientation value for computing any text orientation. These can then be combined to get a single score. The lexicon-based approach can be further categorized into dictionary-based or corpus-based[19]. 3

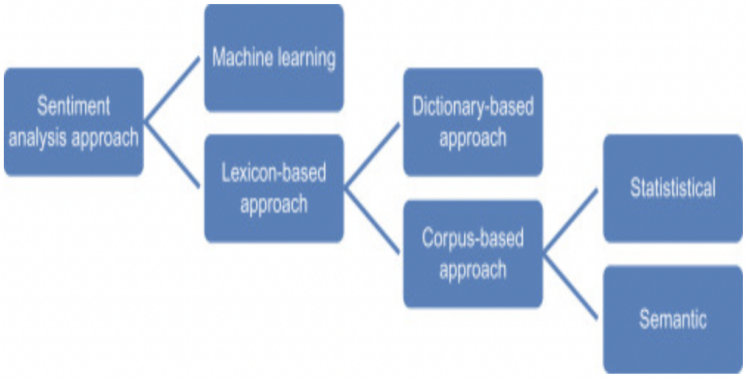

Fig. 3. Lexicon-based

- Dictionary-based method: In this method, a dictionary is built by starting with a few terms. Then, by including synonyms and antonyms of those terms, a thesaurus, WordNet, or an online dictionary can be utilized to build that dictionary. The dictionary gets enlarged until there are no more terms that can be added to it. Manual inspection can help to improve the lexicon.
- Corpus-based method: This method identifies the sentiment polarity of context-sensitive words. The following are the two techniques of this approach:

  Statistical method: Positive polarity is defined as phrases that demonstrate chaotic behavior in positive activity. They have negative polarity if they display negative repetition in the negative text. The term has neutral polarity if the frequency is the same in both positive and negative text.

  Semantic method: This method assigns emotion values to words and words that are semantically similar to those words by locating synonyms and antonyms for the phrase in question.

This method has a number of drawbacks[24]. For example, the existence of more positive terms in any online text source does not always imply that the review is favorable or vice versa. In most circumstances, reusing the same lexicon for scoring texts from distinct domains is impossible. To address this, a new collection of sentiment lexicons, depending on the nature of the particular domain, should be created. By bootstrapping from a smaller vocabulary, some research has been done to construct domain-specific sentiment lexicons for certain target domains.

To lexicon-based approaches, generative models offer a more flexible and alternate answer[16]. This class of models is distinguished by their ability to extract aspects and identify feelings from textual messages at the same time.

2.2 Machine Learning-based sentiment analysis

This approach to sentiment analysis is primarily a form of supervised learning. A set of texts are categorized beforehand by humans. These texts are subjected to text mining and machine learning techniques before their sentiment is predicted. These include data collection, pre-processing, feature extraction, feature selection, and classification[7].

- Data collection: The first step in the machine learning workflow is to collect data for developing the ML model[40]. The accuracy of ML systems' predictions is only as robust as the data used to train them. Some of the issues that can emerge during data collection are as follows[22]:

  Datasets that have been cleansed and are freely available. Take advantage of existing, open-source expertise if the issue statement corresponds with a clean, pre-existing, correctly formed dataset.

  Crawling and scraping the internet. Websites can be crawled and scraped for data using automated tools, bots, and headless browsers.

  Data that is unique to you. For a price, agencies can develop or crowdsource data.
- Data pre-processing: Many errors are likely to be found in social media texts. As a result, when they've been gathered, they're pre-processed into a format that the machine learning model can utilize to build the model[1][20].

  Cleaning up the data: these manual and automatic procedures remove data that has been erroneously added or categorized[43].

  Imbalances in the dataset can be rectified by using methods like bootstrapping, repetition, or the Synthetic Minority Over-Sampling Technique (SMOTE) to generate more observations/samples, which can then be added to the under-represented classes[15].

  Removing punctuation such as .!, @ $() * %

  Removing URLs

  Stopword removal

  Tokenization

  Stemming

  Lower casing
- Feature extraction: This is performed using word embedding techniques which help convert the textual, unstructured data into a numerical form[28]. Different word embedding techniques have different representations in the form of vectors for the same text and thus capture different information. Thus, it is crucial to choose the best word embedding technique for the required task. Extensive research has been carried out to see which word embedding techniques perform better for specific tasks[30].

  Commonly used word embedding techniques include Term Frequency and Inverse Document Frequency (TF-IDF), Continuous Bag of Words (CBOW), Skip-Gram (SKG), Global Vectors for Word Representation (GLOVE), Google news Word to Vector (GW2V), and fasttext (FST)[50]. Bidirectional Encoder Representations from Transformers(BERT) is a relatively new method that has been shown to produce better results[46][17]. 4
- Feature Selection: he performance of the model is dependent on the selection of the relevant features since we use the vectors as inputs. Some techniques used are Analysis of Variance (ANOVA), OneRAttributeEval, GainRatioAttributeEval, ClassifierAttributeEval, Principal Components Analysis, and InfoGainAttributeEval to remove irrelevant features and provide us with the set of relevant and important features[2].

- Classification techniques: The resulting data from the previous ML techniques is provided as input to the classifier, which has been previously trained on the collected data[14]. Many different types of classifiers have been used for sentiment analysis purposes:

  Decision Tree
  Multi-Layer Perceptrons
  K-Nearest Neighbor
  Support Vector Machine
  Artificial Neural Networks/Deep Learning
  Naive Bayes
  Logistic Regression
  Ensemble classifiers such as Random Forest, AdaBoost, etc.

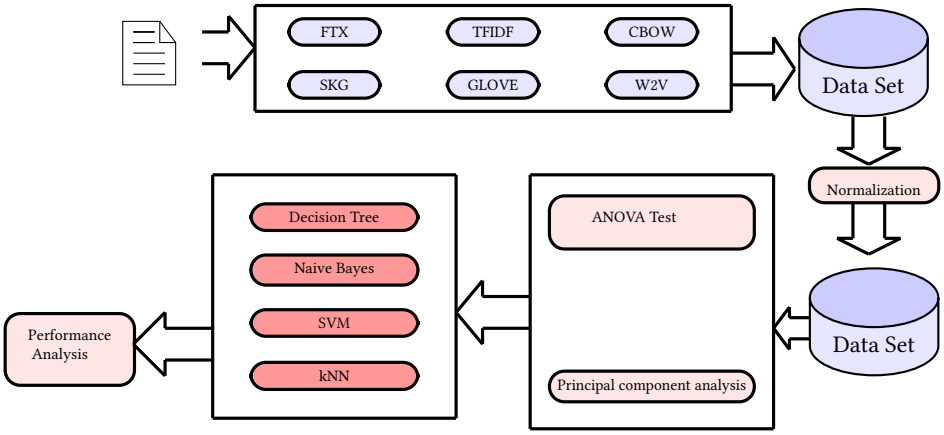

Fig. 4. Research Framework

3 OPINION ANALYSIS

Social media microblogging has become a popular and powerful tool amongst Internet users. Large amounts of information on a variety of topics are shared on these sites in a short amount of time[26]. It might be useful to analyze these microblogs as many informed decisions are made based on these, by the general public. However, traditional opinion mining algorithms do not fit well into social media analysis, given the large volume of information. A micro-blog post is usually very short and colloquial, which does not allow us to use traditional mining algorithms[29].

As online shopping activity is such an important part of current life and continues to grow, the role of online reviews becomes increasingly important[33].

3.1 Aspect Based/Feature-Based Opinion Mining

Aspect-based opinion mining aims to extract aspects and their ratings from a dataset of customer reviews. Not all aspects of a product are reviewed by customers, which makes it essential to summarize the ratings of these aspects to determine the overall review of the product, to allow customers to make informed decisions.

According to C.T.C et al.[10], the core tasks involved in aspect-based opinion mining are:

- Aspect identification

- Aspect based opinion word identification
- Orientation detection

The below figure summarizes the aspect-based opinion mining technique, as described by to C.T.C et al.[10]

5

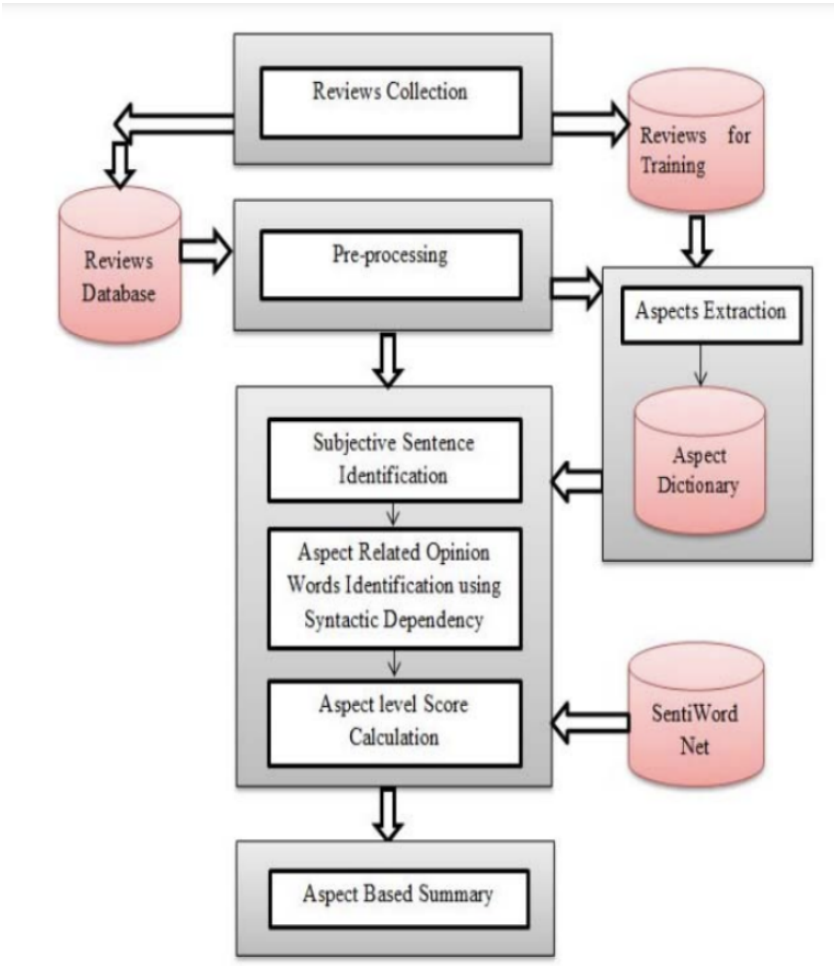

Fig. 5. Aspect-based opinion mining

- Reviews Collection: Parse the HTML webpage containing reviews, and store all your reviews in a database.
- Aspects Extraction: Aspect identification is one of the most complex tasks in opinion mining. It requires the use of Natural Language Processing to extract aspects from reviews.
  Liu[38] uses supervised learning methods(mainly SVM) to find the target aspect for opinion expressions. They create a large corpus of data where each of these are annotated, which they later use to train the SVM. The feature vectors are formed based on the semantic relationships between aspects and their expressions.

One of the initial methods to identify aspects was to treat aspects as nouns and noun phrases and carry out association rule mining algorithms and pruning strategies to find out the candidate aspects[21]. Part-of-Speech parsing was carried out, and a transaction was built from the noun words of each sentence. Then, noun phrases of each statement was fed into an association rule mining model, which returned all frequent itemsets(aspects). The steps are summarized as follows:

Split each review into sentences, and analyse each sentence individually.

Find POS tag of sentence

Extract the words tagged as NN, NNP, NNS, etc.

Count the frequency of extracted words and after removing unnecessary words like 'if', take the most frequent one.

Group synonyms and create an aspect table. Aspect table is a collection of aspects and their synonyms.

- Pre-processing: Before carrying out opinion analysis, we have to pre process the data to remove stop words, non alphabetic characters and emoticons.
- Subjective Sentence Identification: Remove all sentences that do not contain an opinion. Only the sentences with an aspect from the aspect table are considered to be opinionated sentences. The remaining sentences are removed from the database.
- Opinion Words Identification: Opinion words hold opinions towards aspects. These mainly include: adjectives, verbs and adverbs.

  Singh[47] searches n-gram adjacent to the aspect position in a sentence. This method is not suitable if there are reviews for multiple aspects in a sentence.

  Liu[38] uses a Stanford dependency parser to identify opinion words.
- Aspect Score Calculation: This step involves finding the polarity score of the aspect, by using the aggregate of opinion word scores in that sentence. This can be done using tools like SentiWordNet, which is a dictionary of words that assigns scores to sentiment words. The positive or negative score indicates the corresponding polarity of the phrase.
- Calculating Aspect Score In All Reviews: After calculating the aspect score in each sentence, the aggregate positive and negative scores are calculated and after normalization, they are added to find the overall polarity of the aspect.

  We can find the normalized positive and negative scores as follows:

  6 7 Where i refers to the sentence number, and j refers to the aspect number.

$$Normalized\_Positive\_Polarity[j] = \frac{\sum_i Positivepol_{i,j}}{\sum_i}$$

Fig. 6. Equation positive polarity

$$Normalized\_Negative\_Polarity[j] = \frac{\sum_i Negativepol_{i,j}}{\sum_i}$$

Fig. 7. Equation negative polarity

4 TOPIC DETECTION AND TRACKING ON SOCIAL NETWORK

Topic detection and tracking involve the evaluation detecting of a novel, previously unknown topics. Systems must understand what constitutes the topic in specific and thus ensure the independence of topic specifics. The problem revolves around classifying incoming newswire, documents, offline corpus, or broadcast stories into groups containing the same issues. Topic as a summarized tag-set of an input document is differentiable from an event, the latter being a real-world phenomenon with specific spatial and temporal properties[49]. This slight step difference becomes more evident concerning the context of social networks. Detection is termed as identifying current events on media, whereas tracking is jotting down these events and story-boarding. Media here can be single documents, groups of multiple documents, or social media like Twitter, Facebook, and LinkedIn. Concerning social media platforms, detecting events with a vast volume and velocity of data requires extensive research. Twitter, e.g., has expressions that rarely form complete sentences, even contain grammatical errors(intended and unintended) and a bunch of noisiness compared to a news article or a traditional newswire[6]. The data generation factor is also highly impacting. Appropriate information retrieval applications are needed to fit the same. Veracity, correctness, and accuracy of the input media are important data quality. Twitter has an impacting level of fake tweets, rumors, misinformation, or devoid of any information, affecting the tracking systems in work. Tweets being too short, up to 280 characters, unlike traditional newswire, need to be preprocessed to make a long stream of input fit for TDT tasks. Such problems with input media make TDT more extensive, challenging, and complex[31]. 8

4.1 Sub-tasks involved in a general TDT process

The sub-tasks involved in the process have been mentioned below:

(1) Segmentation: separating transitions or setting a fine line distinguishing one topic from another within the document.
(2) Topic Detection: recognition of all topics occurring in the document/matter to be worked upon. It is also termed retrospective topic detection, where cases are available to the input stream, and further grouping needs to be carried out after analyzing each topic[32].
(3) First Story Detection: judging whether the incoming topic already belongs to a known subcategory concerning a specific threshold or needs to be initialized to a new group.
(4) Topic Tracking: classification of incoming new topics into earlier discovered topics; it assumes that a set, $D_c \subseteq D$ of stories belonging to one case $c \in C$ already known by the system, and the incoming stream of topics are to be judged as belonging to the same subject or not. Also known as an unsupervised adaptive tracking task,

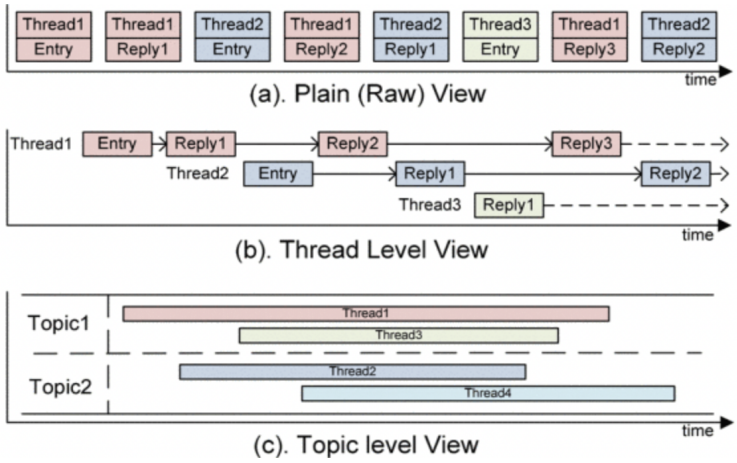

Fig. 8. Different levels of views on threaded discussion communities

the case is also considered when topics regarded by the system as belonging to c are added to D_c. It thus may influence the judgment concerning subsequent topics.

(5) Link Detection: the task of deciding if two given issues belong to the same topic or not.

Support vector machine (SVM) was found to be efficient in training Twitter hashtags metadata when predicting the political alignment of Twitter users. While an incremental online clustering algorithm was used to cluster a stream of Twitter messages in real-time, they also introduced a Naïve Bayes-Text classifier to distinguish between the fastest-growing real-world events and non-events contents on Twitter. The performance of the training set shows the precision of all classifiers computed in the 10-fold cross-validation[12];[9].

9

4.2 TDT on Twitter Network

Dwelling into real-world detection on Twitter Network more precisely[3], the experiments proposed the combination of the following techniques:

(1) LDA (Latent Dirichlet Allocation): Latent Dirichlet Allocation (LDA) extends the generative model to achieve the capacity of generalizing the topic distributions so that the model can be used to generate unseen documents as well[4]. LDA considers the topics to be multinomial distributions over the words and assumes the records to be sampled from a random mixture of these topics. To complete its generative process for the documents, LDA considers Dirichlet priors for the document distributions over issues and the topic distributions over words. The LDA scenario with Twitter has its own issues with tweets being short and less likely to be speaking broadly about a certain issue. T-LDA(Twitter-LDA)also addresses the noisy nature of tweets, where it captures background words in tweets and has outshined LDA. T-LDA has been more specifically also used in aspect mining, and bursty topic detection[39].

(2) Doc-p (Document-Pivot Topic Detection): These approaches cluster together documents using some measure of document similarity, e.g., cosine similarity using a bag of word representation and a TF-IDF weighting scheme.

Detection method	Detection type		Detection task		Data collection Dataset	Detection task
	Event	Topic	RED	NED		
Naïve Bayes classifier		✓		✓	Twitter API, handpicked users	Hot news detection
BScore based BOW clustering	✓			✓	Twitter API (offline)	Disaster and story detection
BOW distance similarity	✓			✓	Twitter API	FSD (first story detection)
BNgram and TF-IDF		✓	✓		Offline datasets	Topic detection
Cross checking via Wikipedia	✓			✓	Twitter API, Wikipedia	Hot news detection
Formal concept analysis		✓		✓	RepLab 2013 dataset	Topic detection
FPM (frequent pattern mining)	✓			✓	Twitter API	Event detection
FPM		✓	✓		Super Tuesday/FA Cup/US elections	Topic detection
FPM (hierarchical clustering)		✓		✓	Topic dataset from CLEar system	Topic detection
FPM (TF-IDF & *n*-gram improved)	✓			✓	Twitter API	Event detection
GPU improved TF-IDF approximation		✓	✓		Offline dataset	Topic detection
BOW similarity	✓			✓	Offline dataset	Topic detection
Word embedding					SemEval dataset	Twitter sentiment classification
Spatiotemporal detection	✓		✓		Offline dataset	Targeted-domain event detection
Clustering of temporal & spatial features	✓		✓		Twitter API	Event detection
Geographical regularity estimation	✓			✓	Twitter API	Geosocial event detection
BOW clustering	✓			✓	Twitter API	Event detection & analysis
Probabilistic modeling	✓			✓	Twitter API	Early disaster detection
FPM	✓		✓		Offline dataset	Event detection
Heartbeat graph	✓		✓		Super Tuesday/FA Cup/US elections	Topic/event detection
Enhanced heartbeat graph	✓		✓		Super Tuesday/FA Cup/US elections	Topic/event detection

Fig. 9. Twitter topic/event detection/tracking related studies

Records are compared using cosine similarity on TF-IDF representations, while a Locality Sensitive Hashing (LSH) scheme is utilized to retrieve the best match rapidly[18].

(3) GFeat-p (Graph-based Feature-Pivot Topic Detection)[34]: This first constructs a graph of sentences (nodes), making sure the textual similarity between two sentences acts as the connection or an edge between them. The saliency of each sentence is computed using a specific centrality measure, e.g., Eigenvector Centrality or its well-known variant, PageRank. It is to be noted that central topics in a cluster are most bound to reflect vital aspects of that specific event than other less major topics. The topic detection problem is considered a graph partition issue and is also solved by a spectral clustering algorithm[53].

(4) FPM (Frequent Pattern Mining)[36]: Involves techniques developed to discover frequent patterns in an extensive database of transactions. In the context of feature-pivot methods, one would look for terms that frequently occur together. FPM has also been used in conjunction with probabilistic topic models to enrich the representation of documents before standard probabilistic topic models process them[25].

5 SOCIAL MEDIA, TEXT MINING AND GRAPH THEORIES

10 Social Media is key to the modern information world, and text mining has a crucial role in analysing data on

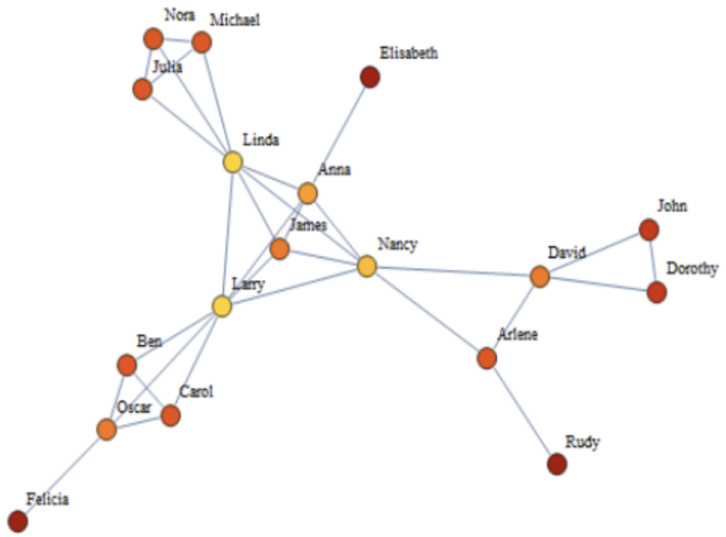

Fig. 10. A network of friends represented by a graph

a day-to-day basis[44] provides us with all the different user-based statistics and the classification of Social Media. Information Extraction, Text Normalisation, Use of Similarity Functions, and Graph Representation are some of the current trends in Text Mining for Social Media.

A Graph, simply put is a bunch of connected objects, called vertices or nodes. The connections could be based on some relation and they're called edges. The concepts of Graph Theory are used extensively in data analysis methods. Coming to social media analysis, the nodes represent users/ set of users and any relations between these nodes whatsoever the reason be, are the edges. Examples of graph theory usage can be seen in Facebook's like, share and tag utilities and Twitter's follow and re-tweet functions to name a few[5].

11

Huan Liu[45] summarized all relevant research subjects in the field of Graph Mining applications for Social Network Analysis, including a roundabout of 60 references from prominent journals and conferences[42]. Talks about the different graph-based representations for text and also look into the Graph-based analysis of text documents for different operations in information retrieval.

The graph-based models work better in representation compared to a regular vector-space model because the meaning of the text is lost and the connection between words cannot be established with the latter.

Simply put, a text document can have a graphical representation in the following ways:

- Co-occurrence graph [37]
- Co-occurrence based on Part-Of-Speech parser [11]
- Semantic graph [11][42]
- Hierarchical Keyword Graph [42]

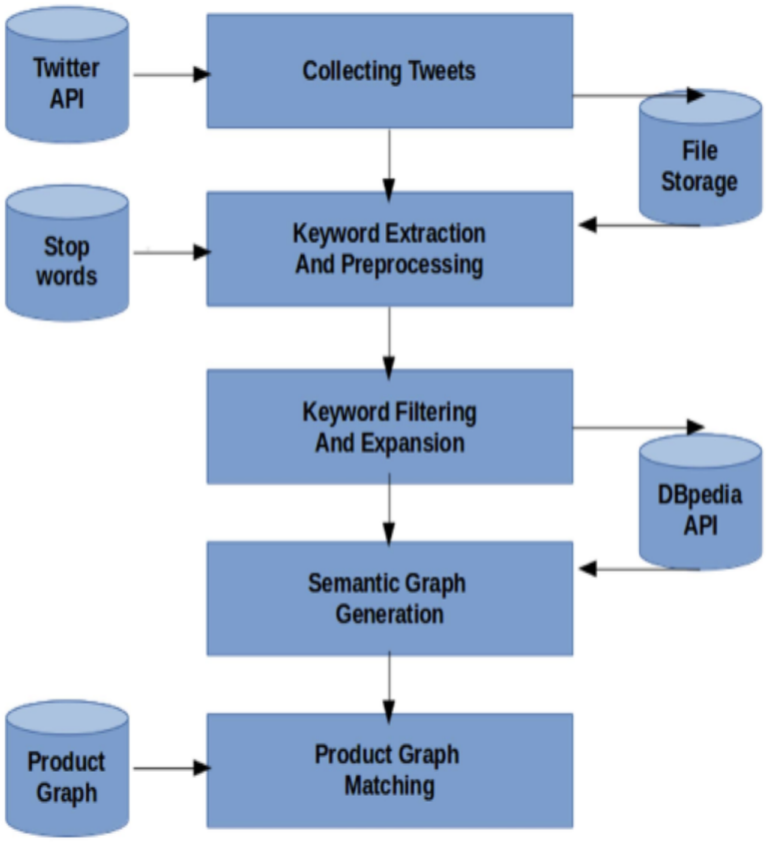

Fig. 11. Feature extraction system

5.1 Social Network Analysis

This involves investigating social media structures using networks and graph theory, i.e in terms of nodes and links. It can have a lot of uses ranging from helping understand one's target audiences to helping businesses take insightful decisions. The key components of Social Media Analytics using Graphs include Density(The "connections" between users) and Centrality(Individual user behavior analysis within a network)[41][27].

5.2 Interest Extraction from user-based texts

Recommender/Rating systems have gained a lot of attention in recent times. Building a good recommender system or personalized systems means to have a detailed knowledge of the personal interests of the user[23]. Jose et al.[23] says this implements a feature extraction system that uses the RAKE algorithm and builds a user graph based on topics extracted from public tweets. The same can be seen in Figure 11.

5.3 Graph-Based Text Classification

One way to efficiently perform text classification with great accuracy is proposed in[48] wherein the method is consuming the term frequency and sentence formation frequency and after the training phase of the algorithm can give accurate orientation for the input text patterns and classified into predefined class labels.

5.3.1 The Proposed Work. This begins with calculating word and sentence probabilities of pre-processed data which are basically the weighted averages. The product of these probabilities gives us what[48] refers to as weight. These weights help us in finding the importance of a word for class orientation an example of which is given in figure 12 [48]. This is then followed by rule extraction and now our model is ready for real-world test datasets.

5.4 Semantic Graph-Based Approach for Text Mining

This approach considers all the nouns of a document and then builds up a semantic graph in a way to represent the document itself. The algorithm for constructing the graph involves normalization, Part-Of-Speech Parser, Noun extractor and word builder[11].

5.5 Cohesive Subgroup Model

This helps in data analysis of clusters of users in large data sets where these "tightly knit" social subgroups are referred to as cohesive subgroups[8]. The model uses k-plex to mine clustered networks on the web.

5.6 Another Graph-clustering Approach

The edges in any graph used for social media analysis can take up multiple forms of relationships, and the way used by Polanco and Juan in[37] is base them on the frequency of co-occurrence of terms indexing the text-data. The concept of "weights" from[48] appears again here because the co-occurrence frequency is not a good enough parameter to measure the strength of these associations on its own.

For this task, we apply the "equivalent coefficient" (as originally defined in Michelet, 1988) based on the product of conditional probabilities of the appearance of a term knowing the presence of the other one.[37]In this type of analysis, the short paths of strong associations can reveal potential new connections between separated fragments of a social framework.

12

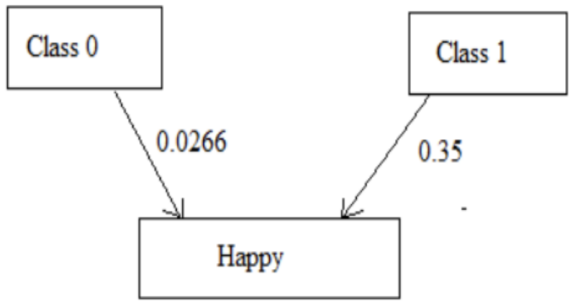

Fig. 12. Text Graph representation for a single word

6 CONCLUSION

Text Mining systems provide us with a plethora of algorithms and tools for personalization and filtering of user information. The main purpose of the survey was to introduce the basic core ideas of commonly used text mining techniques. The premise of our work gets established with the rise of user interactions on social networks, which in turn makes it important for the scientific community to study them more.

This paper starts at the basic definitions of social networks, graphs etc, and moves to analyzing the different algorithms used in text mining - their efficiency and limitations. So far, different levels of successes have been achieved either with singular or combined usage of these techniques.

In this paper, we start by analyzing social networks as graphs with users being the nodes or using words as the nodes. Graph-clustering, graph-based text classification, and interest extraction are some of the things reviewed. We then move to Opinion Analysis where traditional opinion mining algorithms do not work for social media analysis and thus, new approaches were developed to fill the gaps left by the standard ones. The processes involved in Aspect Based/Feature-Based Opinion Mining were discussed in great detail.

The next section is that of Sentiment Analysis which extracts subjective information from a source document. Lexicon-based sentiment analysis which involves estimation of sentiment polarities and Machine Learning based sentiment analysis which includes data collection, pre-processing, feature extraction, feature selection, and classification are discussed. We then study the most interesting part of our work, Topic Detection and Tracking on Social Networks.

Identification of current events on social media and connecting them is not easy given the vast volume of data and the noise that comes with it. Latent Dirichlet Allocation(LDA), Document-Pivot Topic Detection(Doc-p) and Frequent Pattern Mining (FPM) are some of the algorithms studied that help in the real-world detection of Twitter Networks. The present work, as mentioned, can serve as a platform for exploring and developing new methods that can bridge the gaps in the presented approaches.

To conclude, we hope that the reported work on text mining, covering from basic graph-based implementations to practical topic detection methods, can motivate the reader to probe further regarding the diverse conceptual and application-based aspects of text mining.

REFERENCES

[1] Mohammed H. Abd El-Jawad, Rania Hodhod, and Yasser M. K. Omar. 2018. Sentiment Analysis of Social Media Networks Using Machine Learning. In *2018 14th International Computer Engineering Conference (ICENCO)*. 174–176. https://doi.org/10.1109/ICENCO.2018.8636124

[2] Siti Rohaidah Ahmad, Azuraliza Abu Bakar, and Mohd Ridzwan Yaakub. 2015. Metaheuristic algorithms for feature selection in sentiment analysis. In *2015 Science and Information Conference (SAI)*. 222–226. https://doi.org/10.1109/SAI.2015.7237148

[3] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, and Alejandro Jaimes. 2013. Sensing Trending Topics in Twitter. *IEEE Transactions on Multimedia* 15, 6 (2013), 1268–1282. https://doi.org/10.1109/TMM.2013.2265080

[4] Loulwah Alsumait, Daniel Barbara, and Carlotta Domeniconi. 2008. On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking. 3–12. https://doi.org/10.1109/ICDM.2008.140

[5] Sushmita Mondal Anwesha Chakraborty, Trina Dutta and Asoke Nath. 2018. Application of Graph Theory in Social Media. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING* 6 (10 2018), 722–729. https://doi.org/10.26438/ijcse/v6i10.722729

[6] Meysam Asgari-Chenaghlu, Mohammad-Reza Feizi-Derakhshi, Leili Farzinvash, Mohammad-Ali Balafar, and Cina Motamed. 2021. Topic Detection and Tracking Techniques on Twitter: A Systematic Review. *Complexity* 2021 (2021).

[7] EM Badr, Mustafa Abdul Salam, Mahmoud Ali, and Hagar Ahmed. 2019. Social media sentiment analysis using machine learning and optimization techniques. *International Journal of Computer Applications* 975 (2019), 8887.

[8] Balabhaskar Balasundaram. 2008. Cohesive subgroup model for graph-based text mining. In *2008 IEEE International Conference on Automation Science and Engineering*. 989–994. https://doi.org/10.1109/COASE.2008.4626551

[9] Hila Becker, Mor Naaman, and Luis Gravano. 2021. Beyond Trending Topics: Real-World Event Identification on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 5 (Aug. 2021), 438–441. https://ojs.aaai.org/index.php/ICWSM/article/view/14146

[10] Chinsha T C and Shibily Joseph. 2015. A syntactic approach for aspect based opinion mining. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*. 24–31. https://doi.org/10.1109/ICOSC.2015.7050774

[11] Aditi Sharan Chandra Shekhar Yadav and Manju Lata Joshi. 2014. Semantic graph based approach for text mining. In *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*. 596–601. https://doi.org/10.1109/ICICICT.2014.6781348

[12] Michael D. Conover, Bruno Goncalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011. Predicting the Political Alignment of Twitter Users. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 192–199. https://doi.org/10.1109/PASSAT/SocialCom.2011.34

[13] Chedia Dhaoui, Cynthia Webster, and Lay Tan. 2017. Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing* 34 (08 2017), 00–00. https://doi.org/10.1108/JCM-03-2017-2141

[14] Fernando Enríquez, Fermín L. Cruz, F. Javier Ortega, Carlos G. Vallejo, and José A. Troyano. 2013. A comparative study of classifier combination applied to NLP tasks. *Information Fusion* 14, 3 (2013), 255–267. https://doi.org/10.1016/j.inffus.2012.05.001

[15] Wael Etaiwi and Ghazi Naymat. 2017. The Impact of applying Different Preprocessing Steps on Review Spam Detection. *Procedia Computer Science* 113 (2017), 273–279. https://doi.org/10.1016/j.procs.2017.08.368 The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017) / The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2017) / Affiliated Workshops.

[16] E. Fersini. 2017. Chapter 6 - Sentiment Analysis in Social Networks: A Machine Learning Perspective. In *Sentiment Analysis in Social Networks*, Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu (Eds.). Morgan Kaufmann, Boston, 91–111. https://doi.org/10.1016/B978-0-12-804412-4.00006-1

[17] Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. 2019. Target-Dependent Sentiment Classification With BERT. *IEEE Access* 7 (2019), 154290–154299. https://doi.org/10.1109/ACCESS.2019.2946594

[18] Symeon Papadopoulos Georgios Petkos and Yiannis Kompatsiaris. 2014. Two-level Message Clustering for Topic Detection in Twitter. In *SNOW-DC@WWW*.

[19] Neha Gupta and Rashmi Agrawal. 2020. Chapter 1 - Application and techniques of opinion mining. In *Hybrid Computational Intelligence*, Siddhartha Bhattacharyya, Václav Snášel, Deepak Gupta, and Ashish Khanna (Eds.). Academic Press, 1–23. https://doi.org/10.1016/B978-0-12-818699-2.00001-9

[20] Anees Ul Hassan, Jamil Hussain, Musarrat Hussain, Muhammad Sadiq, and Sungyoung Lee. 2017. Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. In *2017 International Conference on Information and Communication Technology Convergence (ICTC)*. 138–140. https://doi.org/10.1109/ICTC.2017.8190959

[21] Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews *(KDD '04)*. Association for Computing Machinery, New York, NY, USA, 168–177. https://doi.org/10.1145/1014052.1014073

[22] Eun Seo Jo and Timnit Gebru. 2020. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 306–316. https://doi.org/10.1145/3351095.3372829

[23] Lijo M. Jose and Rahamathulla K. 2016. A semantic graph based approach on interest extraction from user generated texts in social media. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*. 101–104. https://doi.org/10.1109/SAPIENCE.2016.7684118

[24] S. Kannan, S. Karuppusamy, A. Nedunchezhian, P. Venkateshan, P. Wang, N. Bojja, and A. Kejariwal. 2016. Chapter 3 - Big Data Analytics for Social Media. In *Big Data*, Rajkumar Buyya, Rodrigo N. Calheiros, and Amir Vahid Dastjerdi (Eds.). Morgan Kaufmann, 63–94. https://doi.org/10.1016/B978-0-12-805394-2.00003-9

[25] Hyun Duk Kim, Dae Hoon Park, Yue Lu, and ChengXiang Zhai. 2012. Enriching text representation with frequent pattern mining for probabilistic topic modeling. In *ASIST*.

[26] Peter Kim. 2006. The Forrester Wave: Brand monitoring, Q3 2006. Forrester Wave (white paper).

[27] David Knoke and Song Yang. 2019. *Social network analysis*. SAGE publications.

[28] Yang Li and Tao Yang. 2018. *Word Embedding for Understanding Natural Language: A Survey*. Springer International Publishing, Cham, 83–104. https://doi.org/10.1007/978-3-319-53817-4_4

[29] Po-Wei Liang and Bi-Ru Dai. 2013. Opinion Mining on Social Media Data. *2013 IEEE 14th International Conference on Mobile Data Management* 2 (2013), 91–96.

[30] Yuan Ling, Yuan An, Mengwen Liu, Sadid A. Hasan, Yetian Fan, and Xiaohua Hu. 2017. Integrating extra knowledge into word embedding models for biomedical NLP tasks. In *2017 International Joint Conference on Neural Networks (IJCNN)*. 968–975. https://doi.org/10.1109/IJCNN.2017.7965957

[31] Huailan Liu, Zhiwang Chen, Jie Tang, Yuan Zhou, and Sheng Liu. 2020. Mapping the technology evolution path: a novel model for dynamic topic detection and tracking. *Scientometrics* 125, 3 (2020), 2043–2090.

[32] Zhen Hong Liu, Gong Liang Hu, Tie Hua Zhou, and Ling Wang. 2018. TDT_CC: A Hot Topic Detection and Tracking Algorithm Based on Chain of Causes. In *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. Springer, 27–34.

[33] Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. 2010. Exploiting Social Context for Review Quality Prediction. In *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, North Carolina, USA) *(WWW '10)*. Association for Computing Machinery, New York, NY, USA, 691–700. https://doi.org/10.1145/1772690.1772761

[34] Yiannis Kompatsiaris Manos Schinas, Symeon Papadopoulos and Pericles A. Mitkas. 2018. Event Detection and Retrieval on Social Media. *CoRR* abs/1807.03675 (2018). arXiv:1807.03675 http://arxiv.org/abs/1807.03675

[35] László Nemes and Attila Kiss. 2021. Social media sentiment analysis based on COVID-19. *Journal of Information and Telecommunication* 5, 1 (2021), 1–15. https://doi.org/10.1080/24751839.2020.1790793 arXiv:https://doi.org/10.1080/24751839.2020.1790793

[36] Georgios Petkos, Symeon Papadopoulos, Luca Aiello, Ryan Skraba, and Yiannis Kompatsiaris. 2014. A Soft Frequent Pattern Mining Approach for Textual Topic Detection. In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14) (WIMS '14)*. Association for Computing Machinery, New York, NY, USA, Article 25, 10 pages. https://doi.org/10.1145/2611040.2611068

[37] Xavier Polanco and Eric San Juan. 2006. Text data network analysis using graph approach. In *I International Conference on Multidisciplinary Information Sciences and Technology*, Vol. 2. Open Institute of Knowledge, 586–592.

[38] Bing Liu Qian Liu, Zhiqiang Gao and Yuanlin Zhang. 2013. A Logic Programming Approach to Aspect Extraction in Opinion Mining. *Proceedings - 2013 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2013* 1, 276–283. https://doi.org/10.1109/WI-IAT.2013.40

[39] Minghui Qiu, Feida Zhu, and Jing Jiang. 2013. It is not just what we say, but how we say them: LDA-based behavior-topic model. https://doi.org/10.1137/1.9781611972832.88

[40] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2021. A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. *IEEE Transactions on Knowledge and Data Engineering* 33, 4 (2021), 1328–1347. https://doi.org/10.1109/TKDE.2019.2946162

[41] Olivier Serrat. 2017. *Social Network Analysis.* Springer Singapore, Singapore, 39–43. https://doi.org/10.1007/978-981-10-0983-9_9

[42] Sheetal Sonawane and P. A. Kulkarni. 2014. Graph based Representation and Analysis of Text Document: A Survey of Techniques. *International Journal of Computer Applications* 96 (06 2014), 1–8. https://doi.org/10.5120/16899-6972

[43] Xiaobing Sun, Xiangyue Liu, Jiajun Hu, and Junwu Zhu. 2014. Empirical Studies on the NLP Techniques for Source Code Data Preprocessing. In *Proceedings of the 2014 3rd International Workshop on Evidential Assessment of Software Technologies* (Nanjing, China) *(EAST 2014)*. Association for Computing Machinery, New York, NY, USA, 32–39. https://doi.org/10.1145/2627508.2627514

[44] Triveni Pal Tajinder Singh, Madhu Kumari and Ahsan Chauhan. 2017. Current Trends in Text Mining for Social Media. *International Journal of Grid and Distributed Computing* 10 (06 2017), 11–28. https://doi.org/10.14257/ijgdc.2017.10.6.02

[45] Lei Tang and Huan Liu. 2010. *Graph Mining Applications to Social Network Analysis.* Vol. 40. 487–513. https://doi.org/10.1007/978-1-4419-6045-0_16

[46] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. (2019). https://doi.org/10.48550/ARXIV.1905.05950

[47] Ashraf Uddin V. K. Singh, Rajesh Piryani and Pranav Waila. 2013. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)* (2013), 712–717.

[48] Milap Pathak Vidhyabhushan Dasondi and Narendra Pal Singh. 2016. An implementation of graph based text classification technique for social media. In *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*. 1–7. https://doi.org/10.1109/CDAN.2016.7570879

[49] Guixian Xu, Yueting Meng, Zhan Chen, Xiaoyu Qiu, Changzhi Wang, and Haishen Yao. 2019. Research on topic detection and tracking for online news texts. *IEEE access* 7 (2019), 58407–58418.

[50] Wenkai Yang, Lei Li, Zhiyuan Zhang, Xuancheng Ren, Xu Sun, and Bin He. 2021. Be Careful about Poisoned Word Embeddings: Exploring the Vulnerability of the Embedding Layers in NLP Models. https://doi.org/10.48550/ARXIV.2103.15543

[51] Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. 2019. A Survey of Sentiment Analysis in Social Media. *Knowl. Inf. Syst.* 60, 2 (aug 2019), 617–663. https://doi.org/10.1007/s10115-018-1236-4

[52] Lei Zhang and Bing Liu. 2017. *Sentiment Analysis and Opinion Mining.* Springer US, Boston, MA, 1152–1161. https://doi.org/10.1007/978-1-4899-7687-1_907

[53] Ang Zhao, Xin Lin, and Jing Yang. 2014. Graph-based Model for Topic Detection. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. 1.