# Reviewing Text Mining Techniques for Social Media Analysis

Fanar Fareed Hanna[1], Asst. Prof. Dr. Laith R. Flaih[2]

[1]Salahaddin University-Erbil, Kurdistan Region - Iraq
[2]Department of Computer Science, Cihan University-Erbil, Kurdistan Region - Iraq

***Abstract*—Since the past decade, social media has emerged greatly and become a part of everyday activities of everyone's life. The content being posted on different platforms has increased exponentially and created a huge database, this data has been a gold mine for researchers to dig deep and find trends and patterns, this research aided in detecting spam, removing harmful and offensive content, etc. in this paper the tools and techniques are explained, in addition to the applications that these tools are used for with examples of where they are applied.**

## I. Introduction

Social media platforms have become essential technological tools for information sharing, allowing users to disseminate textual and graphical content. The pervasive use of these platforms, driven by their ease of use and entertainment appeal, has resulted in an unprecedented accumulation of user-generated content [1]. The sheer volume of data shared across diverse social media platforms, encompassing the daily lives and opinions of billions of users, necessitates a systematic analysis [2]. In this context, text mining emerges as a crucial process for extracting meaningful patterns and trends from the vast sea of unstructured data, offering valuable insights that can drive business success [1].

Despite the undeniable importance of text mining in the realm of social media, there exists a need to delve into specific techniques to better understand their nuances and applications. This paper aims to address this gap by exploring various text mining techniques employed in social media platforms, focusing on elucidating the differences among them. By doing so, we aspire to contribute to the broader understanding of how text mining can be harnessed for improved decision-making, strategic planning, and societal insights. To achieve these objectives, this paper is structured as follows:

- Text mining techniques: this section will explain various techniques with examples.

- Data collection: methods used and sources of social media data.

- Challenges: the challenges faced during data collection and mining.

- Applications: where social media text mining is applied and the new evolving trends in social media text mining.

- Tools and software: discussing the popular tools and software used in social media, and evaluation metrics used to assess the performance of text mining techniques.

1- Text mining techniques in social media:

a. Natural language processing (NLP) technique:

Natural language processing gives the ability for computers to understand the spoken and written language of humans, the text that is input to the computer either by microphone (then converted to text) or written is processed by programs that analyze it. NLP is based on two processes, data processing and algorithm development [3].

The NLP technique is important as it plays a key role in text mining, for example, imagine your business software speaks a foreign language that you're not fluent in, the NLP will work as a translator, by taking your human input reorganizing it, and explaining what you've said in a way that your software can parse. The NLP has several techniques that will be tackled in upcoming subsections. [4]

The advantage of NLP can be noticed by the following two examples: "every service-level agreement should include cloud computing insurance," and, "to ensure a well sleep during the night, even if in the clouds, a good SLA to be owned." Suppose a user for searching, relies on NLP. In that case, the algorithm will recognize (cloud computing) as an entity, think of the cloud as an abbreviated form of cloud computing and also take the service-level agreement to be an acronym of SLA. Such elements appear in human language that are usually vague and

DOI:http://doi.org/10.24086/cocos2024/paper.1545

machine learning historically did not well interpret them. Nowadays, with machine learning and deep learning methods, techniques can effectively interpret them [3].

NLP has two main techniques, syntax and semantic analysis. Syntax is a grammatically ordered set of words in a sentence, based on the grammatical rules of a language, NLP uses syntax to extract meanings. The techniques of syntax include:

- Parsing: is the sentence analysis based on language grammar. For example, in the sentence "The bird flew", in parsing the sentence is broken into speech parts, i.e., bird=noun, flew=verb, in complex downstream processing tasks this is most useful.

- Word segmentation: deriving word forms from a string of text, for example, if a person scans a text written page, the algorithm will be able to analyse and recognise the words based on the white space between them.

- Morphological segmentation: dividing the words into smaller parts, these parts are called morphemes, for example, "unforgettable" is broken into [[un][[forget][able]], from which the algorithm will recognize "un", "forget", and "able" as morphemes. This technique is useful in speech recognition and automatic translation.

- Stemming: deriving the word with inflexion to its original root form. For example, in the sentence "The bird flew." the algorithm used will be able to recognize the word "flew" as "fly". [5]

On the other hand, semantics involves using the meaning of the word and behind the meaning, NLP algorithms are used to understand the meaning of the sentence and its structure, semantic techniques include:

- Word sense disambiguation: in this technique, word meaning is derived based on context. For example, the sentence "The pig is in the pen." The word (pen) has various meanings and when using the algorithm of word sense disambiguation, it can comprehend that the word "pen" in this sentence means a fenced area (or barn) and is not a tool used for writing.

- Named entity recognition: this technique determines the categories or groups of words that fall under. For example, consider the sentence "Daniel McDonald's son went to McDonald's and ordered a Happy Meal." This technique will recognize "McDonald's" as two different entities, one is a person and one is a restaurant.

- Natural language generation: this technique uses a database to determine the semantics behind the words and generate new text. For example, generating news based on a body of text used for training, or writing a summary of findings from a BI . [6] [2]

The NLP is used in text extraction, machine translation, and natural language generation, which are applied in a variety of real-world applications, such as:

- Customer-feedback analysis

- Customer-services automation

- Translation

- Academic research and analysis

- Categorization of medical records

- Plagiarism and proofreading and many more applications.

*Sentiment Analysis:*

Is an NLP technique, also called opinion mining, used to determine the existence of sentiment for example, if someone is happy or not happy, and in another example, this technique can determine if the data has positive, negative or neutral feedback [7], as shown in Figure 1

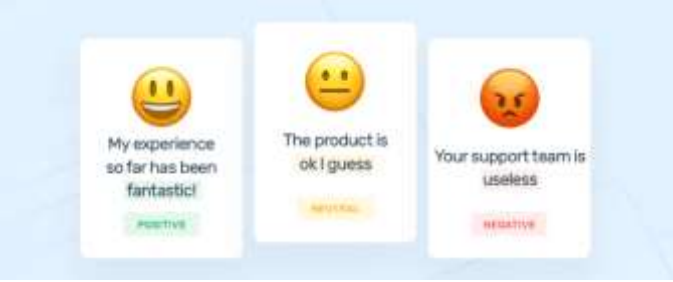


Figure 1. Sentiment Analysis example [8]

In addition to the text polarity which are, positive, negative or neutral, the sentiment analysis technique goes beyond polarity to detect feelings and emotions, urgency and even intentions based on the text. The types of sentiment analysis are usually categorised based on how to interpret customer feedback, below are a set of general examples of sentiment analysis:

- Graded sentiment analysis: it can be considered as the expansion of polarity to have levels of positive and negative, i.e., very positive, positive, neutral, negative, and very negative.

- Emotion detection: the detection of emotions, like anger, happiness, sadness, frustration, etc. uses lexicons (i.e., lists of words and what kind of emotions they convey), but this has one downside of using lexicons, in which word might express emotions differently than what it means, for example, "this bad weather is killing me" versus "Great! You are killing it", the word "kill" in the first sentence is bad, and in the second sentence is expressing happiness.

- Aspect-based sentiment analysis: is used when analysing the features people are rating or evaluating any aspect in a positive,

DOI:http://doi.org/10.24086/cocos2024/paper.1545

neutral or negative way, for example, in this rating of battery life, the customer stated "The battery life of the toy is very short", the aspect-based sentiment will determine that the sentence expresses negative feedback on the battery life.

Social media platforms are becoming the main medium to express almost everything, for example, automatically analyzing the feedback of customers, will help the brands to know what makes their customers happy or frustrated, to improve or change their product to satisfy their customers, the benefits of sentimental analysis can be briefed as: [7]

- Sorting data at a scale
- Real-time analysis
- Consistent criteria

*Topic Modelling:*

Topic modelling is one of the unsupervised machine-learning techniques, which analyses text data automatically to find cluster words from document set. This technique does not need training, of its nature which is unsupervised, hence, it is considered a rapid and effortless way to start data analysis, but with one issue, it cannot guarantee to have accurate results

Topic modelling technique is simple, it consists of counting words and grouping similar word patterns to predict topics within the unstructured data. For example, suppose you want to know what feedback from customers is about a product of a company, and instead of searching through a huge pile of feedback? In that case, the topic modelling technique will help find the clusters of feedbacks that are the same and show the words or expressions that are most often repeated, and is done fast, as this technique does not require training [9].

When comparing topic modelling to topic classification, they only are alike by being mostly used for topic analysis. The topic modelling is unsupervised, hence, less manual input than topic classification which is a supervised technique, however, both require clean and high-quality data and need to be in bucket loads .

At the end of the process topic modelling will produce a collection of documents that are grouped and a cluster of those words or expressions that were used to create the relations. On the other hand, the supervised algorithm (topic classification) will deliver neat and organised results with topic labels or categories such as "Price and UX ", but the issue is they take longer time to set up, as they require to be trained by labelling or marking them with tags the datasets with a ready-made or predefined list of topics, this time expenditure is worth as the results become rewarding with a model that will classify new texts accurately with accordance to their topics. [10]

Topic modelling has several models, such as bag-of-words, unigram model, generative model, etc., the algorithms used for topic modelling are Latent Dirichlet Allocation, Latent Semantic Analysis, Correlated Topic Modelling, and Probabilistic Latent Semantic Analysis. In this paper, we'll talk about LDA and LSA.

*Latent Semantic Analysis:*

LSA is considered the top modelling method used by analysts, is based on the distributional hypothesis that states that the semantics of words can be known from the context the word appears in. In other words, the semantics of a word will be similar if it occurs in a text with a similar context.

The LSA computes the word frequency (word appearance) in the document and the whole collection, it assumes similar documents of context will also have an approximate word distribution and frequency of certain words, in this case, the word order (syntactic information) and multiple meanings of a given word (semantic information) will not be considered and treating each document as a bag of words.

The tf-idf method is used for calculating word frequencies; it also computes in the corpus of all documents how frequent the words are. The words with high frequency will represent the documents. As a result, the representation of tf-idf is considerably better than only consideration of word frequencies at the document level. After computing tf-idf frequencies a document-term matrix states the tf-idf for each term in a given document, for example, Figure 2 shown below.

| Document-Term Matrix | Document 1 | Document 2 | Document 2 | Document 2 |
|---|---|---|---|---|
| Lebron | 0.4 | 0 | 0 | 0 |
| Senate | 0.01 | 0.9 | 0 | 0 |
| Celtics | 0.2 | 0 | 0 | 0 |
| Sprain | 0 | 0 | 0.2 | 0.2 |
| Cancer | 0 | 0.02 | 0.3 | 0.3 |

Figure 2. tf-idf document-term matrix [8]

Three matrices can be derived from the document term matrix (U, S, and V) using SVD (singular value decomposition). The first matrix (U) contains document topics, the (V) matrix contains the terms-topic (example Figure 3), and the linear algebra ensures that S matrix will stay diagonal and LSA will consider each singular value, i.e., every number in the main diagonal of the matrix X is a potential topic to be found in documents.

DOI:http://doi.org/10.24086/cocos2024/paper.1545

Figure 3. USV Matrix [8]

*Latent Dirichlet Allocation:*

The LDA is similar to LSA based on the distributional hypothesis. The goal of using LDA is to map every document in the corpus to a group of topics that cover a vast number of words in the document. The LDA maps the documents to a list of topics by assigning the topics to an arranged list of words, for example, the best player for a topic related to sports. Like LSA, the LDA ignores syntactic information and documents are treated like bags of words, its goal is to decide the topic that the document contains. Figure 4 illustrates how the LDA assumes that the topics and the documents are similar, and when the LDA models a new document, it works as shown in Figure 5.


Figure 4: Assumption of LDA of topics and documents


Figure 5: LDA modelling a document

LSA and LDA have main difference in which the LSA does not assume any distribution which is a vector representation and the LDA the topic distribution is Dirichlet distribution. Two hyperparameters decide the topic and document similarity which are alpha and beta, the value of alpha represents topics for each document and the value of beta will represents the words to model a topic, the lower value of alpha represents less topics of each document and a lower value of beta will have fewer words to shape or model a topic, on the other hand, higher value of any of alpha or beta will have the opposite effect. [10]

### A. *Named Entity Recognition:*

When reading any text from any medium, we immediately identify if this is a person, entity, location, etc., named entity recognition (NER) does this automatically and puts them into predefined categories, as shown in Figure 6 below

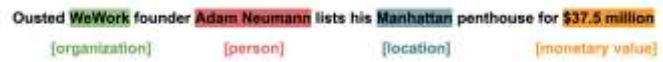


Figure 6: NER recognizing entities [8]

This process is completed through NLP and machine learning, the NLP starts understanding the structure of language and proposes an intelligent system to be capable of deriving meaning from text, based on this system categories of entities are created, a training data is fed to the machine learning model is that will train and improve during time, to be able to classify the entities that exist in the text.

The NER easily identifies the main entities in a text, like locations, brands, people names, etc., with main entities extracted the NER helps sort the unstructured data and extract import information which is crucial when dealing with large datasets. The NER facilitated in many different areas, below is a list of some of these areas: [11]

- Categorize tickets in customer support
- Gain insights from customer feedback
- Content recommendation
- Process CVs

### B. *Text classification:*

Text is an unstructured rich source of information, due to its nature it is difficult to extract information, but thanks to NLP and machine learning both belong to artificial intelligence, sorting and structuring the text for information extraction is getting easier.

This type of classification is considered a machine-learning technique that takes text and categorizes it under a category of predefined categories. For example, organizing IT support tickets by urgency, chat conversations can be organised by language, and so on. Text classification is considered one of the fundamental tools or tasks in NLP that is used in many different applications such as sentiment analysis, topic labelling, spam detection and intent detection.

The importance of text classification can be summarised as follows:

- Scalability: with the help of machine learning vast amount of text can be analyzed in a very small amount of time
- Real-time analysis
- Consistent criteria: humans are prone to mistakes for any reason, but once a model in machine learning is properly trained it will always have the best results.

The classification of text can be achieved in two main ways, manual or automatic. In manual method, as the name indicates, a human can read the text and interpret the meaning to categorize the text accordingly, this method can deliver good results but it is time-consuming and needs a vast number of human powers to perform the task on a huge amount of data. On the other hand, automatic text classification applies machine learning, NLP, and other AI techniques to automatically classify the text at a low cost and faster time. There are three

DOI:http://doi.org/10.24086/cocos2024/paper.1545

main approaches to automatic text classification that all approaches fall under:

- The Rules-based systems
- Machine learning-based systems
- Hybrid systems

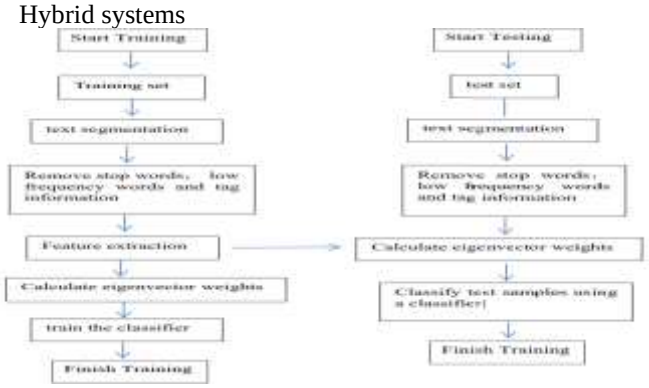


Figure 7: General Process of Text classification

*Rule-based systems:*

The rules-based system is done by creating a set of predefined rules to classify the text, these rules are based on a set of human-created linguistic rules. For example, the task is to categories news into politics and sports, we start by creating a list of words that characterize each category or group, for sports, we include football, basketball, tennis, Ronaldo, Messi, Barcelona, Real Madrid, etc. for political; war, Joe Biden, Oil crises, NATO, etc.

The next step will be taking the input text and counting the words appearing that belong to sports, then doing the same for politics. After that put the text into the category where the highest number of words is shown in that category.

Machine learning-based system:

Creating the rules manually and including all the words could be time-consuming and prone to error especially with a large number of texts, instead using machine learning to make decisions, which is based on past observations.

The process starts with feature extraction, which is a technique used to change the text into numerical form and save it in a vector, in this vector it represents the frequency of a word that is predefined in a dictionary of words. Then, the machine learning algorithm is fed with training data (vectors and their tags) which will teach the algorithm about all words and train it to correctly tag the input text.

Once the algorithm is trained, the testing text is fed to it to predict the category of the text and tag it, Figure 7 is an illustration of the machine learning-based system.

Many algorithms are available for machine learning, the most popular are:

- Naïve Bayes family of algorithms
- Support Vector Machine
- Deep learning

*Hybrid Systems:*

From the name the hybrid system is a combination of both rules and machine learning algorithms, this system is used to enhance the outcome or the results. It can be used to add updated or fine-tuned rules to the rules that have conflicting tags and haven't been modelled correctly. [12]

## II. DATA SOURSE AND COLLECTION

Social media platforms or sites have several categories as shown in (Table 1), accessing these platforms has increased in the last decade with the increase of interest in them, as they facilitate the users the ability to post their photos, messages, and videos. [1]

Data processing involves cleaning and preparing the data to be analyzed by computers, there are several ways to achieve this:

- Tokenization: the text is divided into smaller units (tokens) to work with.

- Stop word removal: includes removing the common words to have the one who offers the most information about the text.

- Lemmatization and stemming: reducing the words to their original forms i.e., their origin or root, to process.

- Part-of-speech tagging: words are labelled depending on the location of part-of-speech they are at, such as nouns, verbs, and adjectives.

After data preprocessing, developing or selecting an algorithm to process it, many options are available but two main types of NLP are commonly used:

- Rules-based system: from the name, it uses linguistics rules that are designed for this purpose, it was implemented early on in the development of NLP and is still used nowadays.

- Machine learning-based system: this type of system is based on statistical methods; it learns based on training data fed to them and performs tasks based on it. Using a combination of neural networks and deep learning, NLP whets their rules through learning from repeated processing. [6]

The usefulness of NLP rises when trying to process massive quantities of unstructured, text-heavy data, most of this data is generated online and stored in databases. Mostly businesses could not process the data effectively until the NLP was introduced.

## III. CHALLENGES IN TEXT MINING FROM SOCIAL MEDIA

Challenges faced by the researchers when collecting text or data from social media include the vast amount of data available that continues to grow in seconds, in addition to the quality and accuracy of the data that can be noisy, incomplete or even fake which can affect the results and decisions built on it.

The social media data is different from the traditional data that we recognize in data mining, apart from the mainly user-generated and noise, it has abundant social relations, such as followers and friendships. This type of data requires special analysis and computational methods to be able to combine social relations theories with statistics and data mining. As mentioned previously in this paper, text-mining techniques

DOI:http://doi.org/10.24086/cocos2024/paper.1545

have an important role in summarizing documents, extracting information, and tagging documents, the same techniques can be used and altered to fit with social media text or data. [13]

Researchers in social media tend to monitor, collect and analyse data to generate or find trends and patterns. Ethical challenges in social analytics exist, and with advances in the technology of analysis using AI, it is advancing over the individual's or organization's awareness of them sharing data with social media platforms. The individual's information being posted on platforms is with their agreement, yet unaware of privacy rules which leads to individual-level issues with privacy breaches, and causes issues to researchers. [14]

## IV. TEXT MINING IN SOCIAL MEDIA APPLICATION

Ther exists a various number text mining applications that are being implemented or used, such as machine learning, Lexicon-based approaches, social text Streams, etc. These are used for text mining and cleansing the data. Real-life applications such as:

### *A. Trend Prediction:*

Trend prediction in text mining of social media plays a vital role, it performs statistical analysis, modelling, and deployment. It is used in many different areas such as, social media platforms, news, and forums to predict the market movement call, for example, up, down and steady. Also, widely used in sentiment analysis.


Figure 4. Figure 8: Trend prediction process in text mining

### *B. Social interaction theories:*

The infrastructure of the exploration mechanism in data mining is the social interaction theory. These theories help us understand online communities and predict who interacts with whom, making them valuable for community detection and link prediction.

### *C. Fake opinion detection:*

One of the crucial activities of data mining on social media is fake opinion detection, the users of social media will post text that has or provides valid information, but rivals tend to post misleading or false information. In e-commerce businesses pay great attention and prioritise feedback and opinions; any false opinion could lead to bankruptcy or the elimination of a product.

One of the known mechanisms to detect and eliminate fake opinions is a Quintuple.

### *D. Spammers:*

It is a known fact that spammers have a big role in causing damage to the truths and social facts on social media platforms, as they will start by creating false news and spread it over the social media platforms, or try to persuade people into making donations, or payments to save others, etc.

Nowadays, the servers or hosts of social media platforms are well equipped with spam detectors that play a vital role in detecting spam text and removing them, for example, removing advertisements with bad language or misleading information.

### *E. Data diffusion:*

Data diffusion machines (DDM) remove unwanted data from social media networks, the work like shared memory (virtually) to spread data across multiple processors. Using a hierarchical directory scheme, they target irrelevant or nonsensical content on social media platforms and remove them. [15]

## CONCLUSION

In this paper, we tackled social media text mining, started by explaining the social media types and what type of data emerges from them, the tools and techniques used to extract information, trends and patterns from it, the challenges faced by researchers when using the data from social media and methods to clean it, in addition to mentioning the software and tools used for the whole procedure, and the tools to measure the accuracy of the models designed for extracting data and information.

## REFERENCES

*[1] N. Al-Kahtani, "Contemporary Emerging Trends of Text Mining Techniques used in Social Media Websites: In-Depth Analysis," in*

DOI:http://doi.org/10.24086/cocos2024/paper.1545

*International Conference on Information Systems and Computer Networks (ISCON), Mahura, India, 2021.*

*[2] F. M. C. J. M. G. José Carlos Cortizo, "Introduction to the Special Issue: Mining Social Media," International Journal of Electronic Commerce,, pp. 5-7, 2011.*

*[3] F. Popowich, "Using text mining and natural language processing for health care claims processing," ACM SIGKDD Explorations Newsletter, pp. 59-66, 1 July 2005.*

*[4] W. M. Marco Varone, "Expert.ai," Expert.AI, 2023. [Online]. Available: https://www.expert.ai/blog/natural-language-processing-and-text-mining/.*

*[5] L. R. a. S. M. H. Flaih, "Software Agent for E-mail Spam Filtering.," International Journal of Applied Engineering Research, vol. 13, no. 10, pp. 7701-7703, 2018.*

*[6] B. Lutkevich, "natural language processing (NLP)," [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP.*

*[7] S. Dessai, "Depression Detection on Social Media Using Text Mining," in 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India., 2022.*

*[8] "MonkeyLearn," MonkeyLearn, [Online]. Available: https://monkeylearn.com/sentiment-analysis/.*

*[9] L. F. Yahya Zakur, "Apriori Algorithm and Hybrid Apriori Algorithm in the Data Mining: A Comprehensive Review," in E3S Web Conf., Online, 2023.*

*[10] J. e. a. Rashid, "Topic Modeling Technique for Text Mining Over Biomedical Text Corpora Through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clustering," IEEE Access, 2019.*

*[11] D. Yanrui, "Named entity recognition method with word position," in IWECAI, 2020.*

*[12] G. Jin, "Application Optimization of NLP System under Deep Learning Technology in Text Semantics and Text Classification," in International Conference on Education, Network and Information Technology (ICENIT), 2022.*

*[13] e. a. Tajinder Singh, "Current Trends in Text Mining for Social Media," International Journal of Grid and Distributed Computing, vol. 10, no. 6, pp. 11-28, 2017.*

*[14] e. a. Elizabeth Ford, "Toward an Ethical Frameowrk for the Text Mining of Social Media for Health Research: A systematic Review," Frontiers in Digital Health, 2021.*

*[15] N. Al-Kahtani, "Contemporary Emerging Trends of Text Mining Techniques used in Social Media Websites: In-Depth Analysis," in 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021.*

DOI:http://doi.org/10.24086/cocos2024/paper.1545