1.1 INTRODUCTION TO DATA MINING

Data mining technology belongs to the analysis of data in order to search for hidden and unforeseen patterns and connections among huge quantity of data. Essentially the purpose of concentration of data mining is to search the hidden facts and unforeseen and transpose it into the meaningful form to utilize in future. Data mining is also named as KDD, knowledge discovery in databases.

Data mining is key process of analyzing data from various sources with different view points and presenting it into a compact manner which is easy to understand and can be helpful in different way. The Data gathered can be utilized to enhance the profits and in cost cutting measures or both. Data miner or data mining tools are basically analytical tools for analyzing the data which can be in various form and the results gathered can be used in multiple way by classifications.

For businesses, data mining is an extremely important platform to discover patterns and relationships among the data in order to help the organizations in decision making. With the help of data mining, businesses can predict sales projections and can help them in making smarter and important decision based on facts and figures.

- **Customer’s taste:** Predict the taste of customers and when they are going to leave the service and the reason behind it.
- **Fraud prediction:** Detection chances in exchange and to make sure whether these are fraud proof.
- **Interactive promoting:** Prediction, in which service detect whether webpages can be of customers preference.
- **Market basket analysis:** It Basically helps to identify and distinguish what kind of products are purchased together like Diapers and Wet wipes or their sales trend during weekdays and during weekends.
- **Trend investigation:** Disclose the distinction between typical clients in- this month and last.
- **Spam detection in Email service:** Predict which email is categorized as spam or ham. Spam emails are illegitimate emails and ham emails are legitimate emails.

1.1.1 THE SCOPE OF DATA MINING

Data mining got its name from searching for exact business information in huge databases. For
example, Flipkart has a huge database of products with each has a unique product identification number and suppose company has a plan to sell those books with older edition or having low sale then rather than manually checking it, mining can be used to draw a pattern because data can be of several Terabytes and exploring it might be a bad decision.

- **Automated prediction of patterns and practices:** Data mining automates the way toward finding prescient data in extensive databases. Questions that generally required extensive hands-on examination can now be addressed specifically from the data rapidly. An ordinary instance of predictive issue is focused on marketing. Data mining uses information on past exceptional mailings to recognize the goals bound to grow rate of productivity in future mailings. Other predictive issues incorporate estimations and different types of default, and recognizing fragments of a populace liable to react comparatively to given occasions.

- **Automated revelation of previously obscure patterns:** Data mining apparatuses cross through databases and recognize effectively disguised examples in a single stage. An instance of example disclosure is the examination of retail bargains information to recognize obviously arbitrary things that are every now and again obtained together. Other example divulgence issues incorporate recognizing false visa trades and perceiving unusual information that could s to information area keying bumbles.

- **Current Scenario:** At the moment, information mining instruments are executed on world class parallel taking care of systems, they can research enormous databases in minutes. Quicker preparing suggests that customers can thusly investigate diverse roads with respect to more models to grasp complex information. Quick makes it practical for customers to dismember enormous measures of information. Greater databases, along these lines, yield upgraded forecasts.

**1.1.2. KNOWLEDGE DISCOVERY PROCESS**

**Knowledge discovery steps are:** Below are the steps to discover knowledge from data.

- **Data cleaning:** The phase of data cleaning is to remove noise, inconsistent and irrelevant data from the database.

- **Data integration:** The process in which dataset are build up by combining multiple data sources.
- **Data selection:** In this phase relevant data is selected from the database according to the analysis.

- **Data transformation:** The phase in which the data are changed into the form that is fit for mining.

- **Data mining:** The procedure where learning strategies are utilized as a part of request to separate out information patterns from data set.

---

![Diagram of Knowledge Discovery Process]

*Fig. 1.1 Knowledge Discovery.*

- **Pattern evaluation:** The step in which patterns is evaluated to recognize the purely attractive patterns that representing knowledge.

- **Knowledge representation:** The step where knowledge representation methods are utilized to demonstrate the understandable knowledge.

### 1.1.3 DATA MINING TECHNIQUE

There are several data mining techniques which have seen deployment in data mining projects. These techniques are given below.
• **Numerical Prediction:** Numerical expectation is the type of directed realizing where we can foresee the numerical esteems, for example, an organization's offers or benefit esteems. Relapse is additionally another name for numerical forecast. One of the prevalent way to deal with numerical forecast is Neural Networks which is a perplexing system in view of human neurons. A neural system is being given distinctive arrangement of sources of info and after that it is utilized to foresee diverse yields.

• **Clustering:** Clustering alludes to clubbing the datasets or items which demonstrate comparative conduct or which have comparable qualities are in one bunch and questions indicating diverse attributes are gathered in various group. Bunching can be portrayed as a technique for autonomous learning and a typical strategy for measurable information investigation which can be helpful in different fields, for example, design acknowledgment, data recuperation or social affair, pictorial examination and so on.

• **Association Rules:** Association rules are the type of unsupervised learning. Association rules are utilized when we need to utilize a preparation dataset to discover the relationship that may exist between the estimations of factors as standards. The common application is “Analysis of Market Basket”. It breaks down the acquiring conduct of the market clients that assistance organization to foresee its future deals. We can likewise characterize it as look for connection between factors.

• **Classification:** Classification is a standout amongst the most imperative utilizations of
information mining. It considers the activity which occurred in our everyday life. For instance, we can group understudies report as legitimacy, pass or fall flat; a doctor's facility may need to classifications its patients into high, low or medium danger of having certain sort of sickness and arrange a text according to the sentiment into either positive and negative. Algorithms of classification are decision tree learning, support vector machine, neural networks etc.

1.1.4 DATA MINING APPLICATIONS

There are different utilizations of information mining. Different ventures have been embracing information digging for their business. The applications of data mining are

- Text mining
- Medical diagnosis
- Pattern mining
- Web mining
- Financial forecasting
- Text summarization
- Weather forecasting
- Trend mining in social networks
- Machine Learning
- Marketing
- Sales
- Manufacturing

1.2 SENTIMENT ANALYSIS

Regular gigantic measure of information is made from interpersonal organizations, online journals and other media and diffused in to the internet. This immense information contains exceptionally essential assessment related data that can be utilized to profit organizations and different parts of business and logical ventures. Manual following and extraction of this helpful data isn't conceivable, accordingly, Sentiment investigation is required. Conclusion Analysis is the wonder of removing suppositions or feelings from surveys communicated by clients over a specific subject, region or item on the web. It is a utilization of normal dialect preparing, computational semantics, and content investigation to distinguish subjective data from source information. It clubs the conclusions in to classes like
"positive" or "negative". Along these lines, it decides the general state of mind of the speaker or an essayist regarding the subject in setting.

Natural language processing (NLP) is the innovation managing our most omnipresent item: human dialect, as it shows up in messages, website pages, tweets, item depictions, daily paper stories, web-based social networking, and logical articles, in a huge number of dialects and assortments. In the previous decade, effective common dialect preparing applications have moved toward becoming piece of our ordinary experience, from spelling and sentence structure rectification in word processors to machine interpretation on the web, from email spam identification to programmed question replying, from distinguishing individuals' conclusions about items or administrations to removing arrangements from your email.

The best test of feeling investigation is to plan application-particular calculations and systems that can examine the human dialect semantics precisely.

Data innovation has advanced throughout the years like no other science. It has contributed in making unthinkable things conceivable. As this examination proposition is being composed, something some place may get developed, changing a few myths, breaking a few taboos; Information innovation assumes its part some place in recording them. Conclusions or feelings as said have been the sole property of human heart and mind which has sounded good to us, people. Data Technology [IT] has taken it over from us and maybe has reshaped our feelings and assessments as well. At this point of mankind's history, innovation can help us in understanding ourselves. Estimation examination is one of the freshly discovered measures that advertisers have found to gauge state of mind of a client via web-based networking media towards a brand or ware or intrigue. The quantities of preferences and takes after on stages like Facebook and Twitter have made a great deal of footing by advertisers in coming to and mining these social stages. These conclusions which are found in input, discussions, or evaluates remarks are filling the Big Data universe with heaps of assorted data that is demonstrated advantageous to different organizations for various causes [Pang and Lee, 2004].

Conversations have remained in the forefront of development and grooming of civilizations ever since human existence has been known. Discussions can be deciphered by various individuals in various ways, however strategies related with Sentiment Analysis help us in understanding the ways the discussion has been hollowed, and what it has implied regardless of varieties in the translating parties. Each discussion passes on some type of assessment, which can extensively be named Positive, Negative and Neutral. A ton of work has been done on feeling examination, significantly in positive and negative naming [Pang and Lee, 2004; Pang et al., 2002; Turney, 2002; Hatzivassiloglou and
As talked about, it is required to comprehend that estimations are negligible emotions, slants and not correct certainties; however, the emotions and slants guide us towards some genuine realities about the general inclination of the feeling transports, the majority, the general population. A supposition or a view is for the most part delegated falling under one of the two contradicting assumption polarities; the feelings/opinions are ordered as parallel in nature – 0/1 or genuine/false or can be comprehended as great/awful, similar to/loathe. This translation is by and large alluded to as **Polarity or Semantic Orientation** [Hatzivassiloglou and McKeown, 1997].

Sentiments Analysis is the investigation of sentiments, demeanors, suppositions and feelings as content. It's a developing and huge zone of research today. These days, individuals utilize different online networking channels like Twitter, Facebook, LinkedIn and so on which gives them a stage to share considerations, encounters, impression with independence.

Sentiment Analysis is a apparatus in information mining that can defeat challenges like outfitting, examining and deciphering printed content since information are scattered, disorder, and divided [Kaplan and Haenlein, 2010], by deliberately separating and investigating on the web information without bringing about whenever delays. With this, advertisers have the chance to find out about customer emotions and states of mind continuously in spite of the difficulties of information structure and volume [Rambocas and Gama, 2013]. These conclusions can be classified either into two classifications: **Positive and Negative** or into an n-point scale i.e. **very good, good, satisfactory, bad, very bad** [Prabowo and Thelwall, 2009]. Sentiment Analysis is used interchangeably with Opinion Mining. It is discussed under the category of Text Mining in concept extraction. Text mining is used to understand the sentiment content of a text unit using various methods such as **Machine Learning, Statistical/Quantitative Techniques or Natural Language Processing (NLP)**. Sentiment analysis can be **Supervised or Unsupervised** [Prabowo and Thelwall, 2009; Abbasi, 2010; Pang and Lee, 2008 (1); Turney, 2002]. Sentiment analysis is used for estimating product quality and accuracy, customer reviews, policy making and political issues [K. Durant and M. Smith, 2006]. Several challenges are focused upon in sentiment classification, **Feature based sentiment classification** and finally **Opinion Summarization** [G.Vinodhini and RM.Chandrasekaran, 2012]. Various levels of sentiment analysis are being supported either at **word level, sentence level or document level** [Anitha N, et al., 2013]. There are many machine learning approaches that are utilized for assumption characterization like max-entropy, data pick up, grouping approaches, support vector machine (SVM) and Naive Bayes (NB). Every one of these methods is utilized for opinion extraction in different types of information.
1.2.1 APPLICATIONS OF SENTIMENT ANALYSIS

Following are the major applications of sentiment analysis in real world scenarios.

- **Product and Service reviews** - The most widely recognized utilization of assessment investigation is in the region of audits of purchaser items and administrations. There are numerous sites that give computerized rundowns of surveys about items and about their particular viewpoints. An outstanding case of that is “Google Product Search”.

- **Reputation Monitoring** - Twitter and Facebook are a point of convergence of numerous supposition investigation applications. The most widely recognized application is observing the notoriety of a particular brand on Twitter or potentially Facebook.

- **Result prediction** - By breaking down slants from pertinent sources, one can anticipate the plausible result of a specific occasion. For example, feeling investigation can give generous incentive to competitors running for different positions. It empowers crusade administrators to track how voters feel about various issues and how they identify with the addresses and activities of the hopefuls.

- **Decision making** - Another critical application is that notion investigation can be utilized as an essential factor helping the basic leadership frameworks. For example, in the money related markets speculation. There are various news things, articles, sites, and tweets about every open organization. An opinion examination system can use these distinctive sources to find articles that discussion about the associations and aggregate the supposition about them as a single score that can be used by a mechanized exchanging framework.

1.3 MACHINE LEARNING

Machine Learning deals with making computers learn to do things the way humans do, primarily it teaches computers to learn the way humans do and for that it makes use of using pattern recognition to make generalizations which are at the core of machine learning. Machine Learning algorithms are evolution over normal algorithms intended to make programs "smarter", by allowing them to automatically learn from the data provided to it.

Machine learning algorithms are provided to program machines and are kept on auto learning mode such that they learn from their experiences and better their performance. Machine Learning (ML)
is a more seasoned region of Artificial Intelligence and takes a shot at overseeing learning through computational techniques. There are numerous cases of gaining from Email informational indexes, natural groupings and more from the web. The Machine learning approach typically banks on training data to generate a model using some features extracted from the training data, in order to make and train the model. The machine learning algorithm employs classification/regression to create the model of correlation between characteristic.

1.3.1 MACHINE LEARNING APPROACHES

Machine learning algorithms are classified into taxonomy, on the basis of the desired outcome of machine learning algorithm. Typical types are detailed as below -

**Supervised learning:** In this category of algorithms, the algorithm generates a function that maps inputs to desired outputs. One of the standard formulations of this category is the classification problem where the learner should learn a function which maps a vector into one of many classes by looking at various Input/Output instances of the function. The system learns to model itself based on the output of the function by iterating on the data subsets. The learning tasks in Supervised Learning [Haykin S, 2009] are as follows:

- Classification - It is used for data/ functions with discrete range
- Regression - It is used for data/ functions with continuous range.

Amongst the various learning methods for supervised machine learning some are Decision Tree Induction, Rule Learning, Linear Regression, Concept Learning, Bayesian Learning, Instance based learning, NN and SVMs. For sentiment labeling, SVM and NN are most used ones, both for classification and regression. SVM are used to reduce the training time and ANN for increasing the accuracy.

**Unsupervised learning:** This category of machine learning models a set of inputs: labeled. This is harder as compared to supervised learning in which the computer learns how to do something that we do not tell it how to do. [Turney, 2002] has given an unsupervised approach to label text as 'Thumbs Up' (recommended) or 'Thumbs Down' (not recommended) [Turney, 2002]. To achieve the same, it follows two approaches - the first approach is to teach the agent by utilizing some kind of reward system to indicate success and not by giving explicit categorizations. The second approach of unsupervised learning is known as Clustering. In this type of learning, the aim is to search similarities in the training data and not to maximize a utility function. The presumption is regularly that the groups
found need to coordinate sensibly well with an instinctive arrangement. For example, bunching people in view of socioeconomics may bring about a bunching of the poor in one gathering and well off in other gathering. The calculation embraces information driven approach that can function admirably when there is adequate information, for example, social data sifting calculations, for example, those that Amazon.com use to suggest books, depend on the standard of finding comparable gatherings of individuals and afterward allotting new clients to gatherings.

*Semi-supervised learning:* This category of machine learning consolidates both marked and unlabeled cases to create a fitting capacity or classifier.

*Reinforcement learning:* In this approach the calculation takes in a strategy of acceptable behavior given a perception of the world. Each activity has some effect in the earth, and nature gives input that aides the learning calculation.

*Transduction:* It is like directed adapting, however it doesn't expressly build a capacity, rather, it tries to anticipate new yields in view of preparing inputs, preparing yields, and new data sources.

*Learning to learn:* This category of algorithm takes in its own inductive predisposition in light of past involvement.

1.3.2 CHALLENGES WITH MACHINE LEARNING FOR SENTIMENT ANALYSIS

A lot of research work has been done in the area of Machine Learning in last few years; however there are still critical issues and challenges with its usage in Sentiment Analysis mentioned as under [Leung, 2006]:

- Data privacy has been sacrificed to deploy benefits of data mining.
- Usage of more powerful computers to apply stronger machine learning technology.
- Use of data for marketing purposes and privacy infringement.
- Machine learning needs to be coordinated into a steady framework building approach as calculations can't give arrangement by working exclusively.

The findings from the experiments performed has provided a means to usage of Machine learning algorithms in Text Categorization [Pang et al., 2002; Turney, 2002]. The two Types of Text Categorization are as follows:

- **Topical Categorization:** In this method of text categorization, documents are sorted on the basis of their subject matter (e.g., society, movies etc.) and are implemented for automation. Typical example can be given as Electronic mail which is flagged with a special sign 'for important'
because of the relevance to people marked in the message. Labeling such messages brings highlight to these messages. Some methods normalize the rating schemes, while few others can filter messages.

- **Non-topic-based text categorization:** In this method of categorization the labeling of text is done on the basis of their source. Every source (e.g. Times Of India vs. Hindustan Times) has its own way of reporting/source style. Therefore though the news item can be the same yet the source style brings variety due to its reporting pattern. This is popularly known as **Genre Categorization** [Pang et al., 2002]. Genre Categorization and subjectivity defection can suggest if the given text contains opinion, but it cannot label the sentiment or determine the nature of the opinion.

1.4 DATA MINING FOR SENTIMENT CLASSIFICATION

Sentiment analysis was first applied to online reviews. Later on when Twitter became popular, [Jansen et al., 2009] realized that microblogging was a potentially rich avenue for companies. From then on, lots of research was based on Twitter. In this chapter pre-processing of the data, the feature selection and different classification algorithms are discussed.

1.4.1 DATA PRE-PROCESSING

Pre-processing in general, is extensively studied and is used in a lot of applications that consider raw, unstructured data. When research became more focused on Twitter, the need of pre-processing increased proportionally, because many tweets are not properly formatted, or consists of mistakes. As [Mike Thelwall et al., 2010] concluded: “Text based communication in English seems to frequently ignore the rules of grammar and spelling”. Therefore, pre-processing techniques are necessary, to acquire a more clean dataset. Cleansing the dataset increases the performance of the later classification system significantly.

In nearly all literature about Twitter sentiment analysis pre-processing techniques used are quite straightforward, like filtering words, letters, punctuations and correcting simple errors.

Correcting the simple type errors, like misspells and repeated letters is based on dictionaries. The dictionaries are used to correct the errors. Similarly, abbreviations and acronyms are replaced with words from a predefined dictionary. Another strategy which has been applied is removing useless content. Since stop words and punctuations don’t have much influence on the sentiment, these are
removed to decrease the diversity of used words in messages. Stop words are extremely common words, which are removed from text in natural language processing in advance. Meanwhile, there is no definite list of stop words. All the dictionaries have therefore to be selected manually from different resources.

An example filtering technique is filtering the overuse of characters. A repetition of vowels, like ’coooooooool’ is an example. An example of repetition of punctuations is ’cool!!!!!!!!’. Those filtering techniques can be realized by recognizing an overuse of more than 2 subsequent characters.

More Twitter specific pre-processing techniques, based on Twitter’s orthography can be used to filter out the special features (hash-tags, user mentions and retweets). Almost all papers that are focused on Twitter filter these kinds of Twitter orthography. With removing usernames, hyperlinks, urls and repeated letters, reduced the number of the features to about 46% of the original. The usernames had the largest impact on the reduction, which is plausible. The diversity of usernames causes a lot of new words, therefore leaving out the usernames results in a great reduction of the data.

The most common technique to replace and filter strings is a regular expression [Ken Thompson, 1968]. It provides a way to filter out errors and overuse, and more generally it provides a decent way to recognize substrings in strings. Regular expressions are a very powerful string matching technique which is used in lot of “search and-replace” functions of applications. They are written in a formal language, which is compact but have the power to match with almost all string patterns.

Another useful technique, called stemming can be applied to reduce the large diversity of words. Stemming is a technique to reduce the conjugations of verbs to their stem, the original root form. An example of stemming is the reduction of”liked” to”lik” and”like” to”lik”. This diminishes the assorted variety of conjugations in which a word can happen, and as outcome lessens the measure of information.

Most of those described techniques are language specific. The language of a text plays an important role in the pre-processing but also in the later classification. As mentioned earlier it is important to acquire a dataset that is as clean as possible. If several languages are considered, they each produce a different structure, grammar and dictionary of words. Therefore, determining the language is necessary to be able to select specific language content. This is the main research area of language identification and has been studied extensively. Papers about sentiment analysis focus solely on one language, most common is the English language, and to a lesser extent the Chinese language. Acquiring a single language dataset requires filtering or translation. Both techniques need to classify the language of the text. Identification of languages can be performed by learning the distribution of characters per
language. In the literature this distribution measure has proven to be effective and simple. The frequency of the letters and subsequent letters define the language. Different implementations use this principle, like TextCat [Cavnar, W. B. and Trenkle, J. M., 1994; LingPipe, 2008]. TextCat is a text identification method, which has been applied to identify the language of newsgroup articles. They achieved outstanding results with an n-gram based approach. Their n-gram approach is based on n contiguous characters, which are individually counted. The k top most n-grams are almost always correlated to the language. The method of n-grams is widely discussed in the next section about feature selection. [LingPipe, 2008] works in a similar way based on the distribution of characters per language.

The outstanding results of 92% and higher from the TextCat language identifier should be refuted, since their data context is different. Their study focuses in particular on well-formatted texts, but tweets are not well-formatted. By first cleansing the tweets, identification may be improved, but will probably not reach the 92%.

After the language of a message has been identified, the model which has to be trained can assume that all messages are from the identified language. Therefore, it can be assumed that only words from the identified language appear in messages.

1.4.2 SUPERVISED CLASSIFICATION

In supervised classification, the system learns to model itself based on the output of the function by iterating on the data with discrete range. There are various techniques for supervised classification as detailed in the sub sections further.

1.4.2.1 FEATURE EXTRACTION

Feature extraction is the technique where properties are separated from the information. These properties, known as features are attributes from the information, in light of the fact that by and by the entire information is too expensive to use in order. The features ought to be discriminative, to portray the first information and additionally conceivable. Then again the highlights ought to lessen the space to avoid redundancies and a high dimensionality of the information. Analysts proposed and tried different things with various highlights in the assessment investigation, to test which extraction techniques functioned admirably. The additional highlights are: n-grams, the tf-idf measure and grammatical form labeling.
N-gram model

The n-gram technique is a generally basic calculation. A n-gram is a coterminal grouping of n things from a printed or talked source. The things are normally letters or words. The letter variation is connected in TextCat and the words variation is connected as highlight extraction technique in this examination. If there should be an occurrence of unigrams (n = 1), every content (tweet) is an archive and is part up into words. Tallying the recurrence of the considerable number of words in an all reports brings about a word recurrence table. High continuous words have a higher likelihood to show up in writings, in this way those words depict the dataset better. There are two possible ways of measuring the frequency among all documents: the summed term frequency and document frequency. Those two frequency measures are based on respectively the term frequency and the term presence. The term frequency \( tf(w,d) \) is the circumstances that a word w belongs to d, stated in equation (1). In computing the term frequency, all occurrences of a word in a document are counted. Therefore, the term frequency can assume a value in the interval \([0−n]\), where n is the aggregate number of words in the archive. The term nearness \( tp(w,d) \) just checks if a word w is available inside a record d, which brings about a twofold esteem (equation (2)). This measure is stated in equation below. The main difference is therewith how the words are counted within a message.

\[
tf(w, d) = |\{w \in d\}|
\]

\[
\begin{align*}
    tp(w, d) &= \begin{cases} 
    1 & \text{if } (w \in d) \\
    0 & \text{if } (w/\in) 
\end{cases} 
\end{align*}
\]

The summed term recurrence \( df(w,D) \) totals all the term frequencies of a word w over all archives \( D \), the formula is stated in equation (3). This is formalized in equation (4), where the document frequency is \( df(w,D) \) of word w across all documents D. Often the set of documents is referred as ‘corpus’ in more formal linguistic jargon.

\[
\begin{align*}
    stf(w,D) &= \sum_{d \in D} tf(w, d) \\
    df(w,D) &= |\{d \in D : w \in d\}|
\end{align*}
\]

Nonetheless, visit words are not really great highlights for order. On the off chance that the conveyance
of an exceptionally visit word is consistently appropriated over the classes, at that point its
discriminative power is low. For example the word "special" is highly frequent but appears as often in
positive, negative and neutral documents, and is therefore not discriminative for one of the classes. In
classification of new cases the classifier can randomly select one of the classes. Therefore a feature
selection method that takes into account how well a feature discriminates between the classes is
additionally used.

Reaching out to higher request n-grams, the writings are not part up as single length things, but rather
n-length things. If there should be an occurrence of n = 2 (bigrams), the things comprise of two back to
back words. The set contains all mixes of two words that are back to back in the first content. Similar
holds for n = 3 (trigrams), and higher n esteems and furthermore for letter based n-grams. Intuitively,
those higher order n-grams seem to capture the relation of the consecutive words better, e.g. in
negation.

Finally, a selection is made of the most valuable words. Only the top k most valuable n-gram features,
according to the weight measure are selected to form the feature vector. This reduces the dimensionality
of the data.

In the literature, several researchers [Charu C. Aggarwal, 2002; Alec Go et al., 2009; Pang et al.,
2002] have implemented classification models based on n-grams. The uni- and bigram variants are
implemented the most, and to a lesser extent trigrams. Comparing those uni- and bigrams does not lead
to an unambiguous answer which is superior. [Pang et al., 2002] accomplished to some degree better
outcomes with unigram term nearness than with unigram term recurrence. They concluded therefore
that the term presence works better than the term frequency. They additionally performed tests with
bigram term nearness and a mix of uni-and bigrams term nearness, yet these outcomes did not beat the
consequences of the unigram term nearness. In the meantime, [Cavnar, W. B. and Trenkle, J. M.,
1994] accomplished better outcomes with the mix of unigrams and bigrams than just unigrams or just
bigrams individually. The separate unigrams still worked out better than the bigrams in his research. In
contrast, [Alexander Pak and Patrick Paroubek, 2010 (1)] achieved better accuracies with the
bigrams than with the unigrams. This could be caused by the different application and the different
domain he used. In accordance with [Pang et al., 2002; Alec Go et al., 2009; Apoorv Agarwal et al.,
2011] tend to the same conclusion that unigrams are better applicable to Twitter than other n-gram
models. It sounds somewhat counterintuitive because bigrams seem to capture the relation between
words better than unigrams. However, the fundamental ideas originate from movie reviews [Pang et
al., 2002] and [Turney, 2002], where models with unigrams outperformed models with higher order n-
TF-IDF Measure
The tf-idf (term frequency inverse document frequency) measure is a measurement that mirrors the significance of a word over an arrangement of records. This measure is made out of two individual measures: the term recurrence and the reverse of the report recurrence. The term frequency and the normal document frequency were discussed in previous sub-section (refer equations (1), (2), (3), (4)).

The converse report recurrence is utilized to quantify the rareness of a word over every one of the archives. How higher the estimation of the reverse archive recurrence, how more uncommon the word over the arrangement of reports is. The converse archive recurrence, idf(w,D) of a word w over all reports D is appeared in equation (5). This can be combined with the term frequency to the tf-idf measure shown in equation (6). In this equation, the tf-idf is the \( tf-idf(w,d,D) \) of a word \( w \) in a document \( d \) across a set of documents \( D \).

\[
idf(w,D) = \log \frac{|D|}{df(w,D)} \tag{5}
\]

\[
tf - idf(w,d,D) = tf(w,d) - idf(w,D) \tag{6}
\]

In any case, it is the issue whether the tf-idf is a decent measure for highlight determination in this examination. The tf-idf esteem is high, when a word happens frequently in a report and does not happen regularly inside all documents. In Twitter domain the documents are the tweets, which mean that words with a high frequency within the tweet and a low frequency over all tweets have a high tf-idf value. These words don't appear to be the best words to use for arrangement, since they do not cover many tweets. Therefore, the tf-idf measure is not considered as feature selection in this research.

Part-of-speech Tagger
A part-of-speech tagger or in no time a POS tagger, is a strategy for increasing a word in a content relating to a specific part of speech. POS is also known as word class or lexical category, which are more intuitive names. Within a text, all words are classified to their corresponding lexical category. The thought behind this is just a constrained arrangement of words in a sentence demonstrate the conclusion, alluded to as the assessment words. In the English dialect cases of lexical classes are thing, verb, qualifier and descriptor. Some of those lexical classes contain assumption words all the more
regularly, for example, modifiers and verb modifiers. A problem of the POS tagger could be a word that can appear in more than one lexical category; however, it only expresses a sentiment in one lexical category.

Since the lexical categories are not the same for all languages, each language needs to use its own POS tagger. The identified language can then be used in the POS tagging method.

The POS labeling method is regularly utilized, which has been connected in a few papers [Apoorv Agarwal et al., 2011; Alec Go et al., 2009; Alexander Pak et al., 2010; Pang et al., 2002; Turney, 2002]. In many papers it is utilized as extra element to a n-gram show, unigrams or bigrams. Go et al. achieved worse results with the POS-tagger than without. Like [Pang et al., 2002; Alec Go et al., 2009] concluded that the POS tagger doesn’t improve the quality of the model. In other papers the authors concluded the same, [Steven Bethard et al., 2006; Efthymios Kouloumpis et al., 2011]. However, there are some papers, e.g. [Alexander Pak et al., 2010], which conclude that a POS-tagger actually improves the performance. They claimed that some POS-tags are good indicators of emotional texts. In line with [Alexander Pak et al., 2010; Apoorv Agarwal et al., 2011] came up with results that improve with the addition of POS features. Additionally, [Luciano Barbosa and Junlan Feng, 2010] achieved better results with a POS-tagger, however his research was based on manually created biased dataset. The results of this last research are therefore less relevant for this research. Since [Turney P D, 2002] didn’t use the POS tagger together with the unigram model, he had no results to compare with. It seems that it is much harder to conclude something about the POS tagger than the n-gram model. There are some researchers which achieved better results with POS-tagging, while others achieved worse results. Even in the literature about movie reviews the POS-tagger did not outperform other models. Because of this discordance, the POS-tagger isn’t a method with convincing properties for sentiment analysis, and is not applied in this research.

1.4.2.2 CLASSIFICATION

Classification, a machine learning technique predicts the class of objects/entities, the class labels of whom are unknown. It involves building a classifier and using that classifier predicting the class labels. It is a learning based approach. The classifier or the model is based on the analysis of training data. Opinion classification is the most important part of the sentiment analysis. It classifies the opinions made in a blog, review or post. It classifies the entities as the positive, negative or neutral. The area of classification algorithms has been studied extensively during the past decades. Classification algorithms like Nearest Neighbor, Naïve Bayes and Support Vector Machine are applied in many
different domains.

**Decision Tree**

A DT classifier uses a tree model to predict the class of an example. The tree consists of one root node, which is where the classifier starts. The other nodes are either leaf nodes, when they have no branches or internal nodes. The internal nodes and the root node represent a feature and a test that has to be performed on that feature. For each possible outcome of the test, the node has a branch that leads to the next node. The leaf nodes eventually indicate a class.

A DT classifier predicts the class of an example by following a path from the root node of the tree until it encounters a leaf node. At every node (except leaf nodes) a test is performed to choose which branch to follow to the next node. When a leaf node is encountered, the classifier predicts the class that the leaf nodes indicate.

Take a look at figure below where a DT is displayed which decides whether to play tennis outside.

Assume the following conditions: Outlook=Sunny and Humidity=High. To decide whether to play tennis, we start at the root node of the tree which is Outlook and follow the branch labelled ‘Sunny’ to the internal node labelled ‘Humidity’. Here we follow the branch labelled ‘High’ to the leaf node labelled ‘no’. This means we will decide not to play tennis under these conditions.

The whole tree could also be represented as if . . . then . . . else . . . rules.

The following rules would represent the same DT:

![Decision Tree Diagram](image)

*Fig. 1. 3 Decision tree.*
IF Outlook = Sunny
    THEN
        { IF Humidity = High
            THEN No
        IF Humidity = Normal
            THEN Yes
    }
IF Outlook = Overcast
    THEN Yes
IF Outlook = Rain
    THEN
        { IF Wind = Strong
            THEN No
        IF Wind = Weak
            THEN Yes
    }

Reduces Error Pruning (REP) Tree Classifier is a quick choice tree learning calculation and depends on the guideline of figuring the data pick up with entropy and limiting the mistake emerging from difference. REP Tree applies relapse tree rationale and creates numerous trees in modified cycles. A while later it picks best one from all generated trees. This calculation develops the relapse/choice tree utilizing fluctuation and data pick up. Additionally, this calculation prunes the tree utilizing diminished blunder pruning with back fitting technique. Toward the start of the model arrangement, it sorts the estimations of numeric properties once. RepTree utilizes the relapse tree rationale and makes various trees in various emphases. After that it chooses best one from all created trees. That will be considered as the agent. In pruning the tree the measure utilized is the mean square mistake on the expectations made by the tree. Essentially Reduced Error Pruning Tree ("REPT") is quick choice tree learning and it manufactures a choice tree in view of the data pick up or lessening the change. REP Tree is a quick choice tree student which fabricates a choice/relapse tree utilizing data pick up as the part standard, and
prunes it utilizing decreased blunder pruning. It just sorts esteems for numeric properties once. Missing esteems are managed utilizing C4.5's technique for utilizing fragmentary examples.

**K- Nearest Neighbor**

K-nearest neighbor classifier is one of the basic directed classifier, which each science student ought to know about. Fix and Hodges proposed K-closest neighbor classifier calculation in the time of 1951 for performing design characterization undertaking.

For straightforwardness, this classifier is called as KNN Classifier. KNN address the example acknowledgment issues and furthermore the best decisions for tending to a portion of the characterization related errands.

The straightforward form of the K-closest neighbor classifier calculations is to foresee the objective name by finding the closest neighbor class. The nearest class will be distinguished utilizing the separation measures like Euclidean separation.

- Compute “d(x, x_i)” i =1, 2, …., n; where d referred to as Euclidean distance among the points.
- Set the computed n Euclidean distances in non-decreasing order.
- Let k be a +ve integer, take the first k distances from this sorted list.
- Search those k-points corresponding to these k-distances.
- Let k_i represents the number of points belonging to the i^{th} class among k points i.e. k ≥ 0
- If k_i > k_j ∀ i ≠ j then put x in class i.

**Nearest Neighbor**

Starting with the simplest of those algorithms, the nearest neighbor algorithm. This could be explained by the ’curse of dimensionality’ in combination with binary variables, like the term presence. Nearest neighbor algorithms have been applied in domains where the number of dimensions is low. It is known from the literature [Charu C. Aggarwal, 2002; Alexander et al., 2000] that nearest neighbor has more difficulties in higher dimensional space. The variance of distances drops and therefore the classification becomes less meaningful. With the binary variables either the distance among points for a specific variable is minimal or maximal, because there is no continuous scale. Contrary to sentiment analysis, in text classification nearest neighbor achieves reasonable results in combination with the tf-idf measure.

The disadvantages of applying the tf-idf measure as feature selection have been discussed above. Therefore the nearest neighbor algorithm is not applied in this research.

**Naive Bayes**

The Bayesian Classification speaks to a managed learning strategy and additionally a factual technique for order. Expect a basic probabilistic model and it enables us to catch vulnerability about the model
principledly by deciding probabilities of the results. It can take care of symptomatic and prescient issues.

The Naïve Bayes is a straightforward probabilistic classifier. It depends on a presumption about shared independency of qualities. Naïve Bayes is based on supervised learning. The goal is to predict the class of the test cases with class information that is provided in the training data. The naïve Bayesian classifier traditionally makes the assumption that a single Gaussian distribution generates numeric attributes.

[Manning et al., 2008] distinguish two naive Bayes’ models in: the multinomial and the Bernoulli. The multinomial naive Bayes model generates one term from the vocabulary in each position of the document. The Bernoulli model generates an indicator for each term of the dataset. The value 1 indicates that a term is present and a 0 indicates absence. This matches with the unigram model that is presented earlier. The Bernoulli model does not count the multiple occurrences of terms whereas multinomial naive Bayes does. Both models assume the same basic principle of Bayes, which is discussed below.

The gullible Bayes calculation utilizes Bayes' hypothesis. The recipe $P(C|F)$ states the contingent likelihood of $C$ given $F$, where $C$ is the class name and $F$ an element.

$$P(C|F) = \frac{p(C)p(F|C)}{p(F)}$$

Bayes hypothesis gives a scientific governs clarifying how you should change your current convictions in the light of new proof. It permits to compute obscure restrictive probabilities from a known contingent likelihood together with the earlier probabilities. In guileless Bayes, it is expected that the factors are free from each other inside each class. In particular, the nearness of a component is random to the nearness of some other feature. In formal language it would be: the posterior probability is registered by increasing the proof scaling factor with the earlier likelihood, duplicated by the result of the free probabilities. The upside of gullible Bayes models is the way that a generally little preparing set is adequate to prepare the model, because independent variables are assumed only the variables for each class are needed. Together with the execution which is frequently very great, the innocent Bayes display is a decent model to use as a kind of perspective for testing the nature of different models.

The Naive Bayes classifier has been connected in a considerable amount of papers about feeling examination. In a more general application in [Pang et al., 2002; Hong Yu and Vasileios
Hatzivassiloglou, 2003] and in Twitter specific papers in [Albert Bifet and Eibe Frank, 2010; GebreKirstos Gebreselassie Gebremeskel, 2011; Alec Go et al., 2009; Alexander Pak and Patrick Paroubek, 2010 (1)], accomplished great outcomes with the multinomial gullible Bayes classifier, it even outflanked the SVM. Another classifier is the most extreme entropy classifier. This classifier has been applied in for example, [Alexander Pak and Patrick Paroubek, 2010 (1); Pang et al., 2002]. Theoretically it should work better than the naive Bayes classifier. However, the opposite is true in many practical problems.

Support Vector Machine

Support vector machine (SVM) is a regulated approach for machine learning. The fundamental thought utilized as a part of SVM is developing a hyper-plane that is ideal for the order of examples that can be directly isolated. This calculation work by plotting every data point in a n-dimensional workspace, where n speaks to the quantity of highlights which are equivalent to the directions in the workspace. The ideal hyper-plane separates the classes.

Support vector machines exist in various structures, straight and non-direct. A help vector machine [Nello Cristianini and John Shawe-Taylor, 2000] is a regulated classifier. What is regular in this unique situation, are the two diverse datasets included with SVM, a preparation and a test set. They assume to have only two features, and therefore a two-dimensional space with a hyper plane. In this two-dimensional case the hyperplane is a line. In the next examples we assume this two-dimensional space for simplification of the explanation. For example, in email spam recognition, the point is to separate the email in two classifies spam or ham email by utilizing an ideal hyper-plane. The thought is to recognize the two classes to accomplish most extreme minor contrast between two classes, viz. spam and ham. SVM speaks to the data focuses in the workspace, mapped with the goal that the data purposes of alternate classify are apportioned by a greatest peripheral contrast. New data indicates are marked that same workspace and expectations are led to examine the class of the new data point. SVM can productively perform non-direct grouping by portion trap (closeness work). Algorithmic strides of SVM for the characterization procedure are as per the following:

- Train the underlying SVM utilizing all the preparation information to have bolster vectors choice capacities.
- Eliminate those help vectors created from the preparation of introductory SVM whose projections have most noteworthy ebbs and flows on the hyper surface by: finding the
projection of the help vectors along the slope of choice capacity utilized, ascertain the thought of shape for each help vector on the hyper-plane, in conclusion sort the help vectors in the diminishing request and deduct the best N-level of the vectors of support.

- Retrain the SVM by left over vectors for best decision.
- Utilize the group of data point to finally prepare the SVM, creating support vectors.

SVM classifiers are grouped into linear and non-linear classifiers, as follows:

**Linear Classifiers:** Separating the data points in linear order by using a hyper-plane is classified as linear classifiers. There are different hyper-plane’s but the best way to separate the data using hyper-plane is by maximum margin difference viz. the distance of hyper-plane and the closest information point of any class. In the perfect circumstance the classes are straightly divisible. In such circumstance a line can be discovered, which parts the two classes splendidly. However, one line parts the dataset splendidly, as well all in all group of lines do. From these lines the best is chosen as the "isolating line". The best line is found by boosting the separation to the closest purposes of the two classes in the preparation set, [Bernhard et al., 1992]. The amplification of this separation can be changed over to an equal minimization issue, which is less demanding to comprehend. The information focuses on the maximal edge lines are known as the help vectors. A case of a linearly separable dataset is presented in figure 1.4.

**Non-Linear Classifiers:** Sometimes the data is not separated properly or linearly in the high dimensional plane for such separation non-linear classifiers are used that correctly classify the information points and label them to their exact class by using kernel tricks. Some mostly used kernel tricks are as follows:

![Fig. 1. 4 A linear SVM with support vectors and a most extreme separation to the two classes. The support vectors are the focuses that lie on the edge, and are underscored with dim blue.](image_url)
• **Homogenous kernels:** Polynomial kernels that are used for analysing the similarity of vectors are represented by the expression below:

\[
k(\overline{a}_i, \overline{a}_j) = (\overline{a}_i \cdot \overline{a}_j)^d
\]

(8)

Where \( k \) is the kernel function and \((\overline{a}_i, \overline{a}_j)\) are the vectors of the work space with \( d \) as the degree of the polynomial.

• **Non-Homogenous kernels:** In Non-homogenous kernels, a free parameter is added that leverage the group of features combined together.

\[
k(a, b) = (a^T b + c)^d
\]

(9)

Practically speaking the classes are typically not directly detachable. In such cases a higher request capacity can part the dataset. To achieve a non-straight SVM classifier an alleged bit trap is connected. A capacity is connected to the dataset which maps the focuses in the non-straight dataset to focuses in a direct dataset. Quite simple possible functions are the square root or the square, which could change the data to a linear space. This computation is done implicitly; therefore, the user does not have to scale the data to a linear space. The only input that is required is which function type with corresponding parameters must be used. Figure 1.5 shows the separating line in a non-linear dataset.

All things considered, practically speaking the model in figure 1.5 isn't exceptionally sensible. Regularly datasets are not pleasantly dispersed with the end goal that the classes can be isolated by a line or higher request work. Genuine datasets contain arbitrary blunders or clamor which make a less spotless dataset. Despite the fact that it is conceivable to make a model that flawlessly isolates the information, it isn't alluring, on the grounds that such models are over fitting on the preparation information. Over fitting is caused by fusing the irregular blunders or clamor in the model. Subsequently the model isn't nonexclusive, and makes altogether more mistakes on different datasets. Making less difficult models shields the model from over fitting. The multifaceted nature of the model must be adjusted between fitting on the preparation information and being bland. This can be accomplished by permitting models which can make blunders. A SVM can make a few blunders to abstain from over fitting. It tries to limit the quantity of mistakes that will be made.
In figure 1.6 an example with outliers is shown. In this example it is not desirable to create a model that perfectly splits the data points. In that case the model is over fitting on the training data, which makes more errors on the test set. The three random outlying data points which are misclassified by the model are shown in red. Sometimes it is impossible to train a model, which achieves a perfect separation. This can only happen when two data points have an identical feature vector and a different class label. A
more mathematical explanation of support vector machines can be found in [Nello Cristianini and John Shawe-Taylor, 2000] or other SVM literature.

Support vector machines classifiers are connected in numerous papers, [Apoorv Agarwal et al., 2011; GebreKirstos Gebreselassie Gebremeskel, 2011; Alec Go et al., 2009; Pang et al., 2002]. They are exceptionally well known in late research. Looking at the Naive Bayes and the SVM classifier, the SVM has been connected the most. [Alec Go et al., 2009] presumed that the three classifiers he utilized, naive Bayes, most extreme entropy and SVM had comparable execution.

In spite of the fact that there exist a few contrasts in the general execution of the classifiers, testing diverse classifiers will build the nature of the framework.

1.4.3 UNSUPERVISED CLASSIFICATION

Point mutual information (PMI) is a straightforward affiliation measure, which can be utilized for unsupervised learning. The characterization depends on the normal semantic introduction, which is expressed beneath. The strategy makes utilization of reference words, for best and most negative affiliation. For the sentiment orientation (SO) a point mutual information measure is used. It quantifies the discrepancy between the probability of the coincidence given the joint distribution and the probability of the coincidence, assuming independence. Point mutual information is defined as:

$$ PMI(w_1, w_2) = \log_2 \left( \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \right) $$

(10)

This PMI measure (equation (10)) is utilized in a sentiment orientation function SO (shown in equation (11)), it formalizes the dependence of the positive and the dependence of the negative sentiment. The reference words ”excellent” and ”poor” are based on the one star and five stars rating respectively in a five star review rating system.

$$ SO(w) = PMI(w, "excellent") - PMI(w, "poor") $$

(11)

This semantic introduction capacity can be connected to all the extricated words in a message. Averaging every one of those assessment introduction estimations of a message brings about a quality. This number can be translated as an estimation as indicated by the inspiration.

The point mutual information measure is used by [Turney, 2002] in which he created an unsupervised
classification algorithm based on the sentiment orientation of phrases of movie reviews. After applying his algorithm, he achieved somewhat disappointing results on a movie database. He explained the bad performance by the difficult movie review dataset. [John Rothfels and Julie Tibshirani, 2010] achieved similar disappointing results with their implementation on reviews.

[Taras Zagibalov and John Carroll, 2008] implemented an almost unsupervised technique to seed word selection for sentiment classification. This seed word selection starts with human-selected seed and uses an iterative method to extract a training sub corpus. Similar to the PMI measure they measured the association with positive and negative words. Their results were much better than with the PMI measures, and in most cases close to the results of supervised classifiers.

The area of unsupervised classification algorithms is somewhat underdeveloped compared to supervised classification algorithms. In the field of Twitter sentiment analysis, unsupervised learning algorithms are even less studied. Whilst reviews mostly have a (star) rating, Twitter doesn’t have a sentiment label, therefore manually labeling is a prerequisite to apply supervised classification algorithms in Twitter domain. Manually labeling data isn’t necessary in the application of unsupervised classification algorithms. However, for the validation of the model a class label is always necessary, for supervised and unsupervised models.

In this research, the applicability of unsupervised classification algorithms will be investigated, but the focus of my research will be more on the supervised algorithms, since they have achieved a better overall performance.

1.5 K-MEANS CLUSTERING

Clustering is one of the very important data mining and machine learning techniques. Clustering is a procedure for discovering groups of closely related elements in the dataset. Many a times the data needs to be clustered into some categories, such as grouping similar users, modeling user behavior, identifying species of Irises, categorizing news items, classifying textual documents, and more. One of the most common clustering methods is K-Means, which is a simple iterative method to partition the data into K – clusters and is the most broadly utilized calculation in the field. The reason for this calculation is to isolate n information focuses into k bunches where the separation between every datum point and its group's middle is limited. At first k-implies picks k arbitrary focuses from the information space, not really focuses in the information, and appoints them as centroids. At that point, every datum point is doled out to the nearest centroid to make k bunches. After this initial step, the centroids are reassigned to limit the separation amongst them and every one of the focuses in their
group. Every datum point is reassigned to the nearest centroid. This procedure proceeds until the point when merging is come to.

The separation between information focuses and centroids can be estimated utilizing a few distinct measurements including the most broadly utilized cosine and Euclidean separations [G. Qian et al., 2004]. Cosine remove is a measure of separation between two information focuses while Euclidean separation is the greatness of the separation between the information focuses. For instance, if two information focuses spoke to two sentences both containing three words in like manner, cosine would give them a separation that is free from how frequently the three words show up in each of the sentences. Euclidean separation, however considers the extent of likeness between the two sentences. “Cosine distance is a measure of distance between two data points while Euclidean distance is the magnitude of the distance between the data points [Daniel Godfrey et al., 2014].

1.5.1 APPLICATIONS OF CLUSTERING

- Clustering examination is extensively utilized as a part of numerous applications, for example, statistical surveying, design acknowledgment, information investigation, and picture preparing.
- Clustering can likewise enable advertisers to find unmistakable gatherings in their client base. What's more, they can describe their client bunches in light of the acquiring designs.
- Clustering additionally helps in determining Taxonomy. In the field of science, it can be utilized to determine plant and creature scientific categorizations, classify qualities with comparative functionalities and pick up knowledge into structures natural to populaces.
- Clustering likewise helps in distinguishing proof of territories of comparable land use in an earth perception database. It likewise helps in the recognizable proof of gatherings of houses in a city as indicated by house write, esteem, and geographic area.
- Clustering additionally helps in grouping reports on the web for data revelation. (Bunch gatherings of clients in view of their entrance designs on WebPages. Group WebPages in view of their substance)
- Clustering is likewise utilized as a part of exception identification applications, for example, discovery of MasterCard misrepresentation.
- As an information mining capacity, bunch investigation fills in as an instrument to pick up understanding into the dispersion of information to watch attributes of each group.
1.5.2 APPROPRIATENESS OF CLUSTERING (K-MEANS) FOR SENTIMENT ANALYSIS

- Scalability – we require exceptionally adaptable grouping calculations to manage substantial databases.
- Ability to deal with different kinds of attributes – we require exceptionally adaptable grouping calculations to manage substantial databases.
- Discovery of clusters with attribute shape – the clustering calculation ought to be fit for recognizing bunches of self-assertive shape. They ought not to be limited to just separation measures that tend to discover round bunch of little sizes.
- High dimensionality – the clustering technique ought to not exclusively have the capacity to deal with low-dimensional information yet in addition the high dimensional space.
- Ability to deal with noisy data – Databases contain uproarious, absent or mistaken information. A few calculations are touchy to such information and may prompt low quality groups.
- Interpretability –The grouping yield ought to be interpretable, conceivable, and usable.

1.6 TEXT ANALYSIS PROCESS

This process involves the following steps:

A) Data Acquisition

In this data acquisition, data are accumulated from various significant sources, for example, web
crawling; twitter tweets, online audit, newsfeeds, report filtering and so forth.

- **Web Crawling**: Web Crawling: It is a methodology for naturally extricating website pages. It is considered a critical part of search engine. By utilizing website page URL address, web crawler finds every one of the pages. Includes following steps:
  - It starts from the page first
  - Read the entire substance of the page
  - Parsing done on pages
  - Repeating the initial step and circulate the methodology until all pages are not captured.
  - Extract the content of crawling pages with the assistance of Html parser.

- **Document Scanning**: It is characterized as a procedure to capture, store, and recover records disregarding their unique arrangement, with the assistance of micrographics and electronic imaging i.e. checking, OCR, ICR, and so forth. Document Scanning or Imaging is the procedure in which original copy are duplicated and saved as computerized symbolism. These pictures are saved as PDF or TIFF files format.

**B) Preprocessing**

It is utilized to remove loud, conflicting and deficient data... Preprocessing includes 3 stages:

- **Tokenization or segmentation**: It is the path toward section a string of formed dialect into its words. Content information contains square of characters known as tokens. So the chronicles are being disengaged as tokens and have been used for additionally taking care of.

- **Removal of stop words**: Expulsion of stop words: Stop words are the words which ought to have been isolated i.e. may be previously or after trademark lingo taking care of. Stop words can't avoid being words which contain insignificant edifying. Diverse gadgets especially avoid to oust these stop words remembering the ultimate objective to help express look for. A couple of aggregations of words can be picked as stop words for any reason. Some web records, ousts by far most of the customary words which consolidate lexical words, for instance, "need" from a substance in order to upgrade execution. It joins English stop words, for instance, "and", "the", "an", "it", "you", "may", "that", "I", "an", "of" et cetera which are considered as 'utilitarian words' as they don't have meaning. Pros have shown that by ousting prevent words from the archive; you can get the upside of diminished record measure without much affecting the exactness of a user's. In any case, thought should be pondered however to take the customer's needs. For the most part, all web crawlers help in taking out the keep words from their records. With the help of wiping out keep words from the document, the record size can be
diminished to around 33% for a word level record. While reviewing the substance of normal dialect handling, which methods for word can be passed on all the more obviously by evacuating the useful word? On the off chance that stop word expulsion is connected, stop list in one content record can't be stacked.

- **Stemming:** Stemming is utilized to depict the procedure to lessen derived words to their starting point word stem. From 60s, the methods for stemming have been thoroughly analyzed in the field of software engineering. For English, the stemmer case are that, it ought to recognize the string "cats", "catty" as in light of the root word "cat", furthermore "strolls", "strolled", "strolling" as taking into account the root word “stroll”.
  
  - **Porter stemmer:** It is characterized as most primitive and best knowing algorithms utilized for stemming.

<table>
<thead>
<tr>
<th>Positive</th>
<th>Negative</th>
<th>Stem using Porter stemmer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common</td>
<td>Commoner</td>
<td>Common</td>
</tr>
<tr>
<td>Desirable</td>
<td>Desire</td>
<td>Desir</td>
</tr>
<tr>
<td>Affection</td>
<td>Affectation</td>
<td>Affect</td>
</tr>
<tr>
<td>Capitalize</td>
<td>Capital</td>
<td>Capit</td>
</tr>
<tr>
<td>Closeness</td>
<td>Close</td>
<td>Close</td>
</tr>
</tbody>
</table>

This should be possible heuristically by recognizing word postfixes i.e. endings and strip them out, with little regularization toward the end.

- **The Lancaster stemmer:** It is a further comprehensively utilized technique for stemming.

<table>
<thead>
<tr>
<th>Positive</th>
<th>Negative</th>
<th>Stem using Lancaster stemmer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arbitration</td>
<td>Arbitrary</td>
<td>Arbit</td>
</tr>
</tbody>
</table>
o **WordNet stemmer:** It has high-accuracy usefulness, yet issue is that it is of restricted use. For affecting changes, it might require a couple of word and grammatical feature tag, where the grammatical feature is considered as adjective - a, noun - n, adverb - r, or verb - v. At the point when such matches are given, it might crumple the tenses, viewpoint, and number checking.

<table>
<thead>
<tr>
<th>Words</th>
<th>Stem using WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>exclaims, verb</td>
<td>Exclaim</td>
</tr>
<tr>
<td>exclaimed, verb</td>
<td>Exclaim</td>
</tr>
<tr>
<td>proved, verb</td>
<td>Prove</td>
</tr>
<tr>
<td>proven, adjective</td>
<td>Proven</td>
</tr>
</tbody>
</table>

**Table 1. 3 Mapping words by WordNet stemmer.**

C) **Data mining**

Applying distinctive mining systems, to infer value about put away data. Diverse mining methodologies are as following: clustering, classification, statistical analysis, natural language processing and so on. In text analytics, for the most part order strategy is utilized. Characterization is a regulated learning strategy that aides in relegating a class mark to an unclassified tuple as per an officially ordered occasion set. Data classifying and identifying is about to tag the data so it can be make rapidly and proficiently.

Be that as it may, different associations can pick up from re-changing their data, which causes keeping in mind the end goal to cut stockpiling and reinforcement costs, with expanding the velocity of data pursuits. Classification can help an association to meet approved and administrative prerequisites to recover particular data inside a particular day and age, and this is most critical
element behind actualizing different data arrangement innovation.

1.7 CLUSTER ANALYSIS

The simple means for analysing and managing large volume and complexity of data is to classify or group the data based on predefined categories or unlabelled clusters items. “Generally, classification technique is either supervised or either it is unsupervised technique solely contingent upon whether they relegate new things to one of a limited number of discrete regulated classes or unsupervised classifications individually [Aditya Joshi et al., 2011; Alex Somla and S.V.N. Vishwanathan, 2009; Anitha N, et al., 2013; Jansen et al., 2009], while clustering is an unsupervised approach. The point of bunching is to segment a limited unlabelled informational index into a limited and discrete arrangement of information structures that are covered up, as opposed to providing exact properties of unobserved samples generated from the same probability of distribution [Aditya Joshi et al., 2011; Andreas M. Kaplan and Michael Haenlein, 2010; Anitha N, et al., 2013] Clustering mainly follows two criteria for partitioning the unlabelled data items that include high similarity between the data objects within the same clusters and low similarity outside the clusters. Clustering leads to the formation of hard clusters.

It is important to learn the difference between classification and clustering. Both these terms can be explained by a simple example. Let a bucket contain some fruits say orange, grapes, apple and banana. Classification works on predefined knowledge or set of information. Classification algorithm will choose any feature like colour to be one of those features and will categorize the fruits depending on that set of information, while clustering has no such model for grouping the objects. Clustering defines its own model say shape to be one of those models. Clustering will group the above fruits based on shape. Clustering process can be carried out [Aditya Joshi et al., 2011; Pang et al., 2002] as follows:-

- **Feature Selection or Extraction:** As referred in [Aditya Joshi et al., 2011; Pang et al., 2002] feature selection is the selection of relevant features from irrelevant features while feature extraction is the production of new features from predefined features.

- **Selection and Designing of Clustering Algorithm:** No such clustering algorithm is available which solves all the clustering problems. Different algorithms are available with different distance formula like Euclidean distance, Manhattan distance, Minkowski distance. It is important to first understand the problem carefully and then decide which algorithm is best suitable.
• **Cluster Validation:** Cluster validation is the assessment of the clusters formed. Testing of the clusters is done to make sure about the quality of the clusters so formed and guarantees that desirable clusters are achieved. Testing of the clusters can be done by three ways external indices, internal indices and relative indices [Aditya Joshi et al., 2011; Akshi Kumar and Teeja Mary, 2012] depending upon the type of clusters.

• **Result Analysis:** The results so formed from the original set of the data is analysed to have an insight view of it and to ensure qualities of clusters formed are satisfied.

### 1.7.1 CLUSTERING METHODS

There are varieties of clustering methods available which are elaborated in this section.

**Partitioning Method**

Assume we are given a database of 'n' objects and the dividing strategy develops 'k' segment of information. Each parcel will speak to a bunch and k ≤ n. It implies that it will characterize the information into k gatherings, which fulfill the accompanying necessity:

- Each gather contains no less than one protest.
- Each question must have a place with precisely one gathering.
- For a given number of parcels (say k), the apportioning technique will make an underlying dividing.
- Then it utilizes the iterative migration procedure to enhance the parceling by moving items from one gathering to other

**Hierarchical Method**

This strategy makes a various leveled disintegration of the given arrangement of information objects. We can group various leveled techniques based on how the progressive deterioration is framed. There are two methodologies here:

- **Agglomerative Approach:** This approach is otherwise called the base up approach. In this, we begin with each question shaping a different gathering. It continues consolidating the items or gatherings that are near each other. It continues doing as such until the point when the greater part of the gatherings is converted into one or until the point when the end condition holds.
- **Divisive Approach:** This approach is otherwise called the best down approach. In this, we begin with the majority of the items in a similar group. In the ceaseless cycle, a bunch is part up into littler groups. It is down until the point that each protest in one group or the end condition
holds. This strategy is inflexible, i.e., once a combining or part is done, it can never be fixed.

**Density-based Method**

This strategy depends on the thought of thickness. The fundamental thought is to keep developing the given group as long as the thickness in the area surpasses some edge, i.e., for every datum point inside a given bunch, the sweep of a given group needs to contain no less than a base number of focuses.

**Grid-based Method**

In this, the articles together frame a lattice. The protest space is quantized into limited number of cells that frame a matrix structure.

### 1.7.2 CLUSTERING BY K-MEANS ALGORITHM

K-Means is the most famous and common type of clustering algorithm used to assign data objects to a cluster. Partitioning algorithm works by dividing the data objects into ‘k’ number of clusters partitions. Let dataset ‘D’, contains ‘n’ number of data items where ‘k’ is the number of partitions, partitioning algorithm assign ‘n’ number of data items to ‘k’ partitioners where (k ≤ n). These ‘k’ partitions are called as Clusters. The data items in a single cluster possess similar characteristics. The number of clusters formed should not be similar to each other as depicted in Fig. 3. Partitioning algorithms ensure that no cluster should be empty [Daniel Godfrey et al., 2014]. Partition algorithm does not follow any hierarchy like hierarchical algorithm follows; it divides the data objects in a single level. Partitioning works efficiently on a large dataset. A common criterion function for generating clustering output in partitioning algorithm is by using squared error algorithm [David Meyer, 2014].

\[ E^2 = \sum_{j=1}^{k} \sum_{i=1}^{n} (\| x_i^{(j)} - c_j \|^2) \]  

(12)

Where \( \| x_i^{(j)} - c_j \|^2 \) is a preferred Euclidean distance, while ‘cj’ are the cluster centres, ‘\( x_i^{(j)} \)’ represent the data objects.

K-Means algorithm is the simplest algorithm which works on iterations to group the data objects in clusters [2]. Following the criteria of the sum of square error, K-Means algorithm has a time complexity of \( O(n) \). Trapping in local optima is the biggest problem with K-Means algorithm if the initial clusters are not chosen with care. K-Means algorithm aims at minimising the squared error. It is an unsupervised learning in which the ‘k’ is fixed apriori defines the initial cluster centre for every cluster the centroid should be defined careful and cunningly because the change in location also leads to change in results. The best way to place the cluster centres is to place them far from each other [16].
After the assignment of centres to the initial clusters, the next step is to associate each nearest point to a particular cluster. When all the data points are placed in the cluster recalculate a new centroid for the clusters. Repeat the process again until no data object changes its location. K-Means algorithm converges faster.

1.7.2.1 ALGORITHMIC STEPS FOR K-MEANS ALGORITHM

Let \( D = (d_1, d_2, d_3, \ldots, d_n) \) be the data points in a dataset and \( S = (s_1, s_2, s_3, \ldots, s_n) \) be the group of centres. The algorithm follows the below-mentioned steps for clustering [16]:

**Step: 1** Manually and randomly select the initial cluster centres ‘c’.

**Step: 2** Use Euclidean distance criterion function and compute the area among all the data points and the clusters centres.

**Step: 3** Information points with minimum distance from cluster centres are placed in that particular cluster.

The figure above (Fig. 1.8) displays the K-means algorithm.

**Step: 4** Above process is carried out for all the data points until all the points are placed in the clusters.

**Step: 5** Recalculate the new cluster focus ‘\( c_i \)’ using equation mentioned below and again calculate the distance of every information point from new cluster centres.

\[
S_i = \frac{1}{c_i} \sum_{j=1}^{c_i} d_i
\]  

(11)

Where, ‘\( c_i \)’ is the number of data points in the \( i^{th} \) cluster

**Step: 6** If no new reassignment of data points takes place then stop the process, if yes repeat the steps 2 and 3.

---

Fig. 1.8 K-Means Algorithm Example.
1.7.2.2 ADVANTAGES OF K-MEANS ALGORITHM

a) K-Means is simple to implement and robust algorithm.
b) K-Means Algorithm converges faster than any other algorithm.
c) The algorithm can work efficiently on large datasets and with a large number of attributes.
d) Tighter and arbitrarily shapes of clusters are formed by K-Means clustering algorithm.
e) Compute faster than Hierarchical clustering algorithm.

1.7.2.3 DISADVANTAGES OF K-MEANS ALGORITHM

a) The algorithm cannot work on categorical data.
b) Trapping into local optima is the biggest disadvantage of K-Means algorithm.
c) Selection of initial clusters and its centroids are done manually and randomly which leads to variation in results if not done carefully.
e) Numbers of initial clusters are fixed apriori