**Sentiment Analysis of Twitter Data: A Survey of**

**Techniques**

**By- Amisha Bartwal**

**Aditi Rawat**

# ABSTRACT

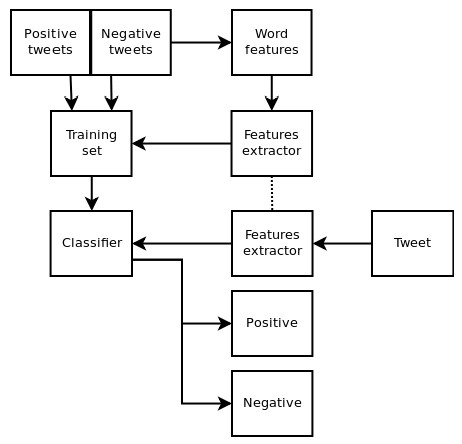
**With the advancement of web technology and its growth, there is a huge volume of data present in the web for internet stoners and a lot of data is generated too. Internet has come a platform for online knowledge, switching ideas and sharing opinions. Social networking spots like Twitter, Facebook, Google are swiftly gaining popularity as they allow people to partake and express their views about motifs, have discussion with different communities, or post dispatches across the world. There has been lot of work in the field of sentiment analysis of twitter data. This check focuses mainly on sentiment analysis of twitter data which is helpful to anatomize the information in the tweets where opinions are largely unstructured, eclectic and are either positive or negative, or neutral in some cases. In this paper, we give a check and a relative analyses of being ways for opinion mining like machine knowledge and dictionary- predicated approaches, together with evaluation criteria. Using various machine learning algorithms like Naive Bayes, Max Entropy, and Support Vector Machine, we give disquisition on twitterdatastreams.We have also mooted general challenges and operations of Sentiment Analysis on Twitter.**

# Keywords 1. Twitter, Sentiment analysis( SA), Opinion mining, Machine knowledge, Naive Bayes( NB), Maximum Entropy, Support Vector Machine( SVM). 2. prolusion presently, the age of Internet has changed the way people express their views, opinions. It's now mainly done through blog posts, online forums, product review websites, social media,etc. presently, millions of people are using social network spots like Facebook, Twitter, Google Plus,etc. to express their passions, opinion and share views about their quotidian lives. Through the online communities, we get an interactive media where consumers inform and impact others through forums. Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, commentary, reviews,etc. also, social media provides an occasion for businesses by giving a platform to connect with their guests for advertising. People mainly depend upon user generated content over online to a great extent for decision timber.Fore.g.However, also they firstly look up its reviews online, bandy about it on social media before taking a decision, if someone wants to buy a product or wants to use any service. The amount of content generated by stoners is too vast for a normal user to anatomize. So there is a need to automate this, various sentiment analysis ways are considerably used. Sentiment analysis( SA) tells user whether the information about the product is satisfactory or not before they buy it. Marketers and enterprises use this analysis data to understand about their products or services in such a way that it can be offered as per the user ‟ s conditions. Textual Information recovery ways mainly concentrate on processing, searching or assaying the factual data present. Data have an objective element but, there are some other textual contents which express private characteristics. These contents are mainly opinions, sentiments, appraisals, stations, and passions, which form the core of Sentiment Analysis( SA). It offers multitudinous challenging openings to develop new operations, mainly due to the huge growth of available information on online sources like blogs and social networks. For illustration, recommendations of particulars proposed by a recommendation system can be predicted by taking into account considerations analogous as positive or negative opinions about those particulars by making use of SA.

# SENTIMENT ANALYSIS

Sentiment analysis can be defined as a process that automates mining of stations, opinions, views and passions from text, speech, tweets and database sources through Natural Language Processing( NLP). Sentiment analysis involves classifying opinions in text into orders like" positive" or" negative" or" neutral". It's also appertained as subjectivity analysis, opinion mining, and appraisal birth. The words opinion, sentiment, view and belief are used interchangeably but there are differences between them. • Opinion A conclusion open to disagreement( because different experts have different opinions) • View private opinion • Belief deliberate acceptance and intellectual assent • Sentiment opinion representing one „ s heartstrings An illustration for languages for Sentiment Analysis is as given below, = The story of the movie was weak and boring = = = = = Sentiment Analysis is a term that include multitudinous tasks analogous as sentiment birth, sentiment type, subjectivity type, summarization of opinions or opinion spam discovery, among others. It aims to anatomize people's sentiments,, stations, opinions passions,etc. towards rudiments analogous as, products, individualities, motifs, associations, and services. Mathematically we can represent an opinion as a quintuple o, f, so, h, t), where o = object; f = point of the object o; so = exposure or opposition of the opinion on point f of object o; h = opinion holder; t = time when the opinion is expressed. ObjectAn reality which can be a, person, event, product, association, or content FeatureAn particularity( or a part) of the object with respect to which evaluation is made. Opinion exposure or polarityThe exposure of an opinion on a point f represent whether the opinion is positive, negative or neutral. Opinion holder The holder of an opinion is the person or association or an entitythat expresses the opinion. In recent times a lot of work has been done in the field of “ Sentiment Analysis on Twitter “ by number of researchers. In its early stage it was intended for double type which assigns opinions or reviews to bipolar classes analogous as positive or negative only. Pak and Paroubek( 2010)( 1) proposed a model to classify the tweets as objective, positive and negative. They created a twitter corpus by collecting tweets using Twitter API and automatically annotating those tweets using emoticons. Using that corpus, a sentiment classifier predicated on the multinomial Naive Bayes system that uses features like N gram and POS- labels. The training set they used was less effective since it contains only tweets having emoticons. Parikh and Movassate( 2009)( 2) executed two models, a Naive Bayes bigram model and a Maximum Entropy model to classify tweets. They set up that the Naive Bayes classifiers worked much better than the Maximum Entropy model. GoandL.Huang( 2009)( 3) proposed a result for sentiment analysis for twitter data by using distant supervision, in which their training data comported of tweets with emoticons which served as noisy labels. They make models using Naive Bayes, MaxEnt and Support Vector Machines( SVM). Their point space comported of unigrams, bigrams and POS. They concluded that SVM outperformed other models and that unigram were more effective as features. Barbosa etal.( 2010)( 4) designed a two phase automatic sentiment analysis system for classifying tweets. They classified tweets as objective or private and also in alternate phase, the private tweets were classified as positive or negative. The point space used included retweets, hashtags, link, punctuation and exclamation marks in convergence with features like former opposition of words and POS. Bifet and Frank( 2010)( 5) used Twitter streaming data handed by Firehouse API, which gave all dispatches from every user which are privately available in real- time. They experimented multinomial naive Bayes, stochastic grade descent, and the Hoeffding tree. They arrived at a conclusion that SGD- predicated model, when used with an applicable knowledge rate was the better than the rest used. Agarwal etal.( 2011)( 6) developed a 3- way model for classifying sentiment into positive, negative and neutral classes. They experimented with models analogous as unigram model, a point predicated model and a tree kernel predicated model. For tree kernel predicated model they represented tweets asatree.The point predicated model uses 100 features and the unigram model uses over, 000 features. They arrived on a conclusion that features which combine former opposition of words with their corridor of- speech( pos) labels are most important and plays a major rolein the type task. The tree kernel predicated model outperformed the other two models. Davidovetal.,( 2010)( 7) proposed a approach to use Twitter user- defined hastags in tweets as a type of sentiment type using punctuation, single words, n- grams and patterns as different point types, which are also combined into a single point vector for sentiment type. They made use of K- Nearest Neighbor strategy to assign sentiment labels by constructing a point vector for each illustration in the training and test set. Po- WeiLianget.al.( 2014)( 8) used Twitter API to collect twitter data. Their training data falls in three different orders( camera, movie, mobile). The data is labeled as positive, negative andnon- opinions. Tweets containing opinions were filtered. Unigram Naive Bayes model was executed and the Naive Bayes simplifying independence supposition was employed. They also barred useless features by using the collaborative Information and Chi square point birth system. ultimately, the exposure of an tweet is predicted. i.e. positive or negative. Pabloet.al.( 9) presented variations of Naive Bayes classifiers for detecting opposition of English tweets. Two different variants of Naive Bayes classifiers were erected videlicet birth( trained to classify tweets as positive, negative and neutral), and Binary( makes use of a opposition dictionary and classifies as positive and negative. Neutral tweets neglected). The features considered by classifiers were Lemmas( nouns, verbs, adjectives and adverbs), opposition wordbooks, and Multiword from different sources and Valence Shifters. Turney et al( 11) used bag- of- words system for sentiment analysis in which the connections between words was not at each considered and a document is represented as just a collection of words. To determine the sentiment for the whole document, sentiments of every word was determined and those values are united with some aggregation functions. Kamps etal.( 12) used the verbal database WordNet to determine the emotional content of a word along different confines. They developed a distance metric on WordNet and determined semantic opposition of adjectives. Xia etal.( 13) used an ensemble frame for Sentiment type which is attained by combining various point sets and type ways. In thier work, they used two types of point sets( Part- of- speech information and Wordrelations) and three base classifiers( Naive Bayes, Maximum Entropy and Support Vector Machines). They applied ensemble approaches like fixed combination, weighted combination and Meta- classifier combination for sentiment type and attained better delicacy.Luoet.al.( 14) stressed the challenges and an effective ways to mine opinions from Twitter tweets. Spam and madly varying language makes opinion recovery within Twitter challenging task.

A General model for sentiment analysis is as follows,



## Fig.1. Sentiment Analysis Architecture

Following are the phases required for sentiment analysis of twitter data,

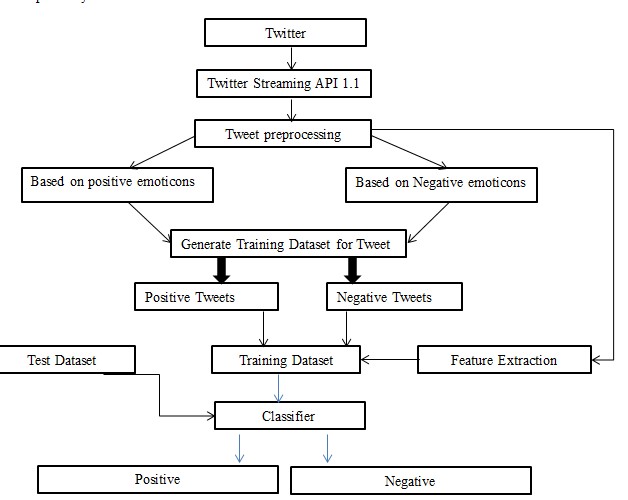
## Pre-processing of the datasets

### A tweet contains a lot of opinions about the data which are expressed in different ways by different druggies. The twitter dataset used in this check work is formerly tanford into two classesviz. negative and positive opposition and therefore the sentiment analysis of the data becomes easy to observe the effect of colorful features. The raw data having opposition is largely susceptible to inconsistency and redundancy. Preprocessing of tweet include following points, Remove all URLs(e.g.www.xyz.com), hash markers(e.g.#topic), targets(@username) Correct the spellings; sequence of repeated characters is to be handled Replace all the emoticons with their sentiment. Remove all punctuations, symbols, figures Remove Stop Words Expand Acronyms( we can use a acronym wordbook) · RemoveNon-English Tweets

### Table 1. Publicly Available Datasets For Twitter

|  |  |  |  |
| --- | --- | --- | --- |
| HASH | Tweets | <http://demeter>.inf .ed.ac.uk | 31,861 Pos tweets 64,850  Neg tweets,  125,859 Neu tweets |
| EMOT | Tweets and  Emoticons | <http://twittersenti>  ment.appspot.co m | 230,811 Pos&  150,570 Neg tweets |
| ISIEVE | Tweets | [www.i](http://www.i)-sieve.com | 1,520 Pos tweets,200 Neg  tweets, 2,295  Neu tweets |
| Columbia univ.dataset | Tweets | Email:  apoorv@cs.colum bia.edu | 11,875 tweets |
| Patient dataset | Opinions | <http://patientopin> ion.org.uk | 2000 patient opinions |
| Sample | Tweets | <http://goo>.gl/Uqv dx | 667 tweets |
| Stanford dataset | Movie Reviews | <http://ai>.stanford.  edu/~amaas/data/ sentiment/ | 50000 movie reviews |
| Stanford | Tweets | <http://cs>.stanford. edu/people/alecm  go/trainingandtest data.zip | 4 million tweets  categorized as  positive and negative |
| Spam  dataset | Spam Reviews | <http://myleott>.co m/op\_spam | 400 deceptive and 400 truthful  reviews in  positive and negative category. |
| Soe dataset | Sarcasm and nasty reviews | http://nlds.soe.uc sc.edu/iac | 1,000 discussions, ~390,000 posts, and some ~  73,000,000 words |

## Feature Extraction

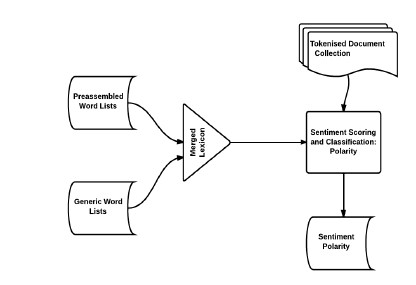
* 1. The preprocessed dataset has multitudinous disparate groupings. In the point birth system, we prize the aspects from the reused dataset. subsequently this aspect are exercised to reckon the positive and inhospitable opposition in a judgment which is useful for arbitrating the opinion of the identities utilizing models like unigram, bigram( 18). 2. Engine knowledge ways bear defining the pivotal features of text or documents for processing. These pivotal features are c o n s i d e r e d as point vectors which are exercised for the type task. Some samples features that have been reported in literature are 3. Words And Their frequence Unigrams, bigrams and n- gram models with their frequency censuses are considered as features. There has been farther disquisition on utilizing word presence preferably than frequence to further describe this point. Pan getal.( 23) showed off better effects by utilizing presence preferably of frequence. 4. corridor Of Speech markers corridor of peroration like adjectives, adverbs and some groups of verbs and nouns are good hands of subjectivity and passion. We can produce syntactic dependence patterns by parsing or dependence trees. 5. Opinion Words And Expressions gradually from special words, some expressions and expressions which convey sentiments can be exercised asfeatures.e.g. cost someone an arm and leg. 6. situation Of tours The situation of a tenure with in a text can affect on how much the tenure makes disparity in common passion of the text. 7. Contradiction Contradiction is an important but delicate point to interpret. The presence of a contradiction usually changes the opposition of the opinion., I am not happy. 6. Syntax Syntactic patterns like tours are exercised as features to get subjectivity patterns by multitudinous of the researchers. Training headed knowledge is an important fashion for working out type cases. Training the classifier makes it easier for future predictions for unknown data. Bracket Naive Bayes It's a probabilistic classifier and can get the pattern of examining a set of documents that has been allotted( 9). It compares the contents with the list of words to codify the documents to their right order or class. Allow d be the tweet and c \* be a class that is assigned to d, where C arg mac
* • Pc NB( c d|) m P c()) p f(| c) ni d() PNB( c| d) i 1 P d() From the below equation, „ f ‟ is a „ point ‟, count of point( fi) is denoted with ni( d) and is present-day in d which represents a tweet. also, m denotes no. of features. Parameters P( c) and P( f| c) are reckoned through ultimate incommodity evaluations, and smoothing is assumed for unseen features. To train and codify utilizing Naïve Bayes Machine Learning fashion, we can exercise the Python NLTK archive. ultimate Entropy In Maximum Entropy Classifier, no hypotheticals are taken descrying the relationship in between the features pulled from dataset. This classifier invariably tries to maximize the entropy of the system by estimating the conditional division of the class marker. ultimate entropy indeed handles overlap point and is same as logistic regression system which finds the division over classes. The conditional division is outlined as MaxEnt makes no independence hypotheticals for its features, unlike Naive Bayes. The model is described by the following exp( i f c di(,)) PME( c d|,) i c exp( i f c di(,)) i Where c is the class, d is the tweet and λ is the cargo vector. The cargo vectors decide the significance of a point in type. Brace Vector Machine 3. Brace vector engine analyzes the data, outline the resolution boundaries and uses the kernels for computation which are performed in input room( 15). The input data are two sets of vectors of size m each. also every data which described as a vector is codified into a class. Nextly we detect a fringe between the two classes that is far from any document. The distance defines the fringe of the classifier, maximizing the fringe reduces indecisive opinions. SVM also supports type and regression which are useful for statistical knowledge proposition and it also helps recognizing the procurators exactly, that needs to be taken into account, to understand it successfully. 4. APPROACHES FOR passion dissection There are mainly two ways for passion dissection for the twitter data Engine mastering Approaches Engine knowledge predicated path uses type fashion to codify text into classes. There are mainly two manners of engine knowledge ways. Unsupervised knowledge It does not correspond of a order and they do not give with the accurate targets at all and therefore calculate on clustering.. headed knowledge It's predicated on labeled dataset and thus the labels are handed over to the model during the process. These labeled dataset are trained to get meaningful labors when encountered during resolution- forestland. The success of both this knowledge styles is mainly depends on the election and birth of the special set of features exercised to detect passion. The engine knowledge path workable to passion dissection mainly belongs to supervised type. In a engine knowledge ways, two sets of data are demanded. Training Set 2. Test Set. A number of engine knowledge ways have been formulated to codify the tweets into classes. Engine knowledge ways like Naive Bayes( NB), ultimate entropy( ME), and brace vector motors( SVM) have achieved great success in passion dissection. Engine knowledge starts with collecting training dataset. Nextly we train a classifier on the training data. Once a supervised type fashion is named, an important resolution to make is to handpick point. They can tell us how documents are described. The most usually exercised features in passion type are • tenure presence and their frequency • portion of peroration information • Negations • Opinion words and expressions 

### Fig.2 Sentiment Classification Based On Emoticons

## With reference to supervised ways, brace vector motors( SVM), Naive Bayes, Maximum Entropy are some of the most common or garden ways exercised. Whereassemi-supervised and unsupervised ways are proffered when it isn't practicable to have an original set of labeled documents opinions to codify the rest of particulars

## Lexicon-Based Approaches

Wordbook grounded system( 20) uses passion wordbook with opinion words and match them with the data to determine opposition. They assigns passion grudges to the opinion words describing how Positive, inhospitable and Objective the words contained in the wordbook are. wordbook- grounded approaches substantially calculate on a passion wordbook, i.e., a collection of known and precompiled passion tours, expressions and indeed expressions, developed for traditional stripes of message, similar as the Opinion Finder wordbook;



**Fig 3.Lexicon-Based Model**

There are Two sub classifications for this approach:

*3.2.1.Dictionary-based:*

It's grounded on the operation of tours( seeds) that are generally collected and annotated manually. This set grows by searching the antonyms and synonyms of a wordbook. An illustration of that wordbook is WordNet, which is exercised to develop a gloss called SentiWordNet. Debit Ca n’t deal with sphere and environment special exposures. . Corpus- Grounded The corpus- grounded path have ideal of furnishing wordbooks related to a special sphere. These wordbooks are generated from a set of seed opinion tours that grows through the hunt of affiliated words by means of the use of either statistical or semantic ways. styles grounded on statistics idle Semantic dissection( LSA). styles grounded on semantic similar as the use of antonyms and synonyms or connections from gloss like WordNet may also represent an intriguing result. tallying to the interpretation measures like perfection and recall, we give a relative study of being ways for opinion mining, involving engine literacy, wordbook- grounded approaches, trial sphere andcross-lingual approaches,etc., as shown off in Table 2.

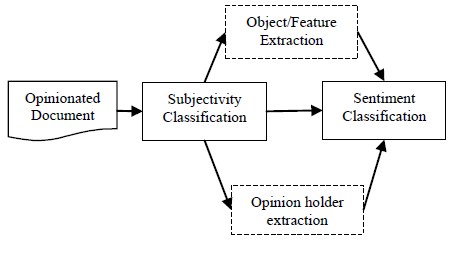
**Table 2. Performance Comparison Of Sentiment Analysis Methods**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Method** | **Data Set** | **Acc.** | **Author** |
| Machine  Learning | SVM | Movie reviews | 86.40% | Pang,  Lee[23] |
| CoTraining  SVM | Twitter | 82.52% | Liu[14] |
| Deep learning | Stanford Sentimen  t  Treebank | 80.70% | Richard[18] |
| Lexical based | Corpus | Product reviews | 74.00% | Turkey |
| Dictionary | Amazon‟  s  Mechani cal Turk | --- | Taboada[20] |
| Crosslingual | Ensemble | Amazon | 81.00% | Wan,X[16] |
| Co-Train | Amazon,  ITI68 | 81.30% | Wan,X.[16] |
| EWGA | IMDb movie review | >90% | Abbasi,A. |
| CLMM | MPQA,N  TCIR,ISI | 83.02% | Mengi |
| Crossdomain | Active Learning | Book, DVD,  Electroni cs,  Kitchen | 80%  (avg) | Li, S |
| Thesaurus | Bollegala[22  ] |
| SFA | Pan S J[15] |

# SENTIMENT ANALYSIS TASKS

Sentiment analysis is a challenging interdisciplinary task which includes natural language processing, web mining and machine learning. It is a complex task and can be decomposed into following tasks, viz:

* Subjectivity Classification
* Sentiment Classification
* Complimentary Tasks
* ObjectHolderExtraction
* Object/Feature Extraction



## Fig.4 Sentiment Analysis Tasks

1. **Subjectivity classification**
2. Subjectivity bracket is the task of categorizing rulings as opinioned or not opinionated. Allow S = { s1,..., sn} be a set of rulings in document D. The case of subjectivity bracket is to identify rulings exercised to represent opinions and other forms of subjectivity( private rulings set Ss) from rulings exercised to objectively present factual information( existential rulings set consequently), where SsUSo = S.
3. **Sentiment Classification**
4. Once the task of chancing whether a judgment is opinionated is done, we've to detect the opposition of the judgment i.e., whether it expresses a positive or inhospitable opinion. passion bracket can be a double bracket( positive or inhospitable),multi-class bracket( extremely inhospitable, inhospitable, neutral, positive or extremely positive), retrogression or ranking. Depending upon the operation of passion dissection, subtasks of opinion proprietor birth and object point birth can be treated as voluntary.
5. **Complimentary Tasks**
   * *OpinionHolder Extraction*

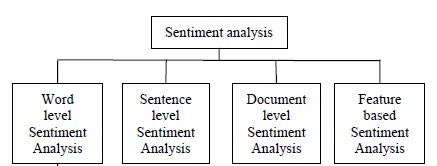
It is the discovery of opinion holders or sources. Detection of opinion holder is to recognize direct or indirect sources of opinion.

* + *Object /Feature Extraction*

It is the discovery of the target entity.

# LEVELS OF SENTIMENT ANALYSIS

Tasks described in the previous section can be done at several levels of granularity.



**Fig.5 Levels Of Sentiment Analysis**

## Document level

* It deals with tagging individual documents with their passion. In Document position the entire document is codify eitherinto positive or inhospitable class. General Approach detect the passion oppositeness of individual rulings or words and combine them together to detect the opposition of the document. Other approaches Establishment verbal marvels likeco-reference conclusion, pragmatics,etc. colorful Tasks involved in this are • Task passion Bracket of entire document • Classes Positive, inhospitable and neutral
* Assumption :Each Document focuses on a single object (not true in discussion posts, blogs, etc.) and contain opinion from a single opinion holder

## Sentence or phrase level

Judgment- position passion dissection deals with tagging individual rulings with their separate passion oppositeness. judgment position passion bracket classifies judgment into positive, inhospitable or neutral class. General path detect the passion exposure of individual words in the judgment / expression and also to combine them to determine the passion of the entire judgment or expression. Other approaches call converse structure of the textbook colorful Tasks involved in this are • Task 1 relating private/ ideal rulings Classes ideal and private • Task 2 passion Bracket of rulings Classes positive and inhospitable Assumption A judgment contains only one Opinion which may not invariably be true Aspect position or point position It deals with labeling each word with their passion and also relating the reality towards which the passion is directed. Aspect or point position passion bracket enterprises with relating and rooting product features from the source data. ways like reliance parser and converse structures are exercised in this. colorful Tasks involved in this are • Task1 Identify and prize object features that have been reflected on by an opinion proprietor( eg. A critic) • Task2 arbitrating whether the opinions on features are inhospitable, positive or neutral • Task 3 detect point antonyms Word Level Most recentworks have exercised the previous opposition of words and expressions for passion bracket at judgment and document situations Word passion bracket exercise substantially adjectives as features but adverbs, The two styles of automatically annotating passion at the word position are 1) Dictionary- Grounded Approaches 2) Corpus- Grounded Approaches. 4. EVALUATION OF passion Bracket The interpretation of passion bracket can be estimated by utilizing four indicators calculated as the following equations delicacy = ( TP TN)( TP TN FP FN) Precision = TP/( TP FP) Recall = TP/( TP FN) F1 = ( 2 × Precision × Recall)( Precision Recall) In which TP, FN, FP and TN relate independently to the number of true positive cases, the number of false negativeinstances, the number of false positive cases and the number of true inhospitable cases, as outlined in the table 1. **Table 3. Confusion Matrix**

|  |  |  |
| --- | --- | --- |
|  | Predicted Positives | Predicted Negatives |
| Actual Positive | TP | FN |
| Actual Negative | FP | TN |

# RESULTS AND DISCUSSION

We used the twitter dataset publicly made available by Stanford university. Analyses was done on this labeled datasets using various feature extraction technique. We used the framework where the preprocessor is applied to the raw sentences which make it more appropriate to understand. Further, the different machine learning techniques trains the dataset with feature vectors and then the semantic analysis offers a large set of synonyms and similarity which provides the polarity of the content.

**Dataset Description:**

|  |  |
| --- | --- |
| **Train Data** | 45000 |
| **Negative** | 23514 |
| **Positive** | 21486 |

|  |  |
| --- | --- |
| **Test Data** | 44832 |
| **Negative** | 22606 |
| **Positive** | 22226 |

**A. Baseline Algorithm:**

The baseline algorithm used is Naïve Bayeswithout preprocessed data and unigram model. Following table shows the accuracy obtained at different sizes for the baseline algorithm.

## Table 4. Accuracy of Baseline Algorithm

|  |  |
| --- | --- |
| **Dataset** | **Accuracy** |
| 10 | 0.46475731620 |
| 50 | 0.533324411135 |
| 100 | 0.54744379015 |
| 500 | 0.612375089222 |
| 1000 | 0.652301927195 |
| 5000 | 0.697403640257 |
| 10000 | 0.712928265525 |
| 15000 | 0.717389364739 |
| 20000 | 0.722764989293 |
| 25000 | 0.729478943612 |
| 30000 | 0.729122055675 |
| 35000 | 0.73244557459 |
| 40000 | 0.733226266952 |
| 45000 | 0.736549785867 |

Following are the details on most informative features after the classifier is executed on train data.

|  |  |
| --- | --- |
| sad = True | neg : pos = 37.6 : 1.0 |
| worst. = True | neg :pos = 32.4 : 1.0 |
| crying = True | neg : pos = 24.7 : 1.0 |
| fml = True | neg : pos = 24.1 : 1.0 |
| hurts = True | neg : pos = 21.2 : 1.0 |
| awful = True | neg : pos = 21.1 : 1.0 |
| ugh. = True | neg :pos = 20.4 : 1.0 |
| terrible = True | neg : pos = 20.4 : 1.0 |
| boo. = True | neg :pos = 19.2 : 1.0 |

cancelled = True neg : pos = 19.2 : 1.0

**B. Naïve Bayes Algorithm:**

**Effect of Stop words**

When Naïve Bayes(Baseline)wasrun,itgaveanaccuracyof73.65percent, which is considered as the baseline result .Then used as removal of Stop word .When stop words were removed and Naive Bayes was run, I tgaveanaccuracyof74.56percent.Following table shows the accuracy obtained at different sizes for the Naïve Bayes with stop words removed and using pre processed data and based on unigram model.

**Table 5. Accuracy of Naïve Bayes Algorithm (Stopword removal +unigram)**

|  |  |
| --- | --- |
| **Dataset** | **Accuracy** |
| **10** | 0.522305496074 |
| **50** | 0.583333333333 |
| **100** | 0.593839221984 |
| **500** | 0.649134546752 |
| **1000** | 0.673536759458 |
| **5000** | 0.7005710207 |
| **10000** | 0.717300142755 |
| **15000** | 0.725486259814 |
| **20000** | 0.731441827266 |
| **25000** | 0.734653818701 |
| **30000** | 0.738891862955 |
| **35000** | 0.740743219129 |
| **40000** | 0.742148465382 |
| **45000** | 0.745605817273 |

Most Informative Features For Naïve Bayes with stopwords removed and unigram model are, bummed = True neg : pos = 34.8 : 1.0 disappointed = True neg : pos = 28.8 : 1.0 sad = True neg : pos = 27.6 : 1.0 awful = True neg : pos = 20.3 : 1.0 ugh = True neg : pos = 19.3 : 1.0 poor = True neg : pos = 19.3 : 1.0 sucks = True neg : pos = 18.7 : 1.0 upset = True neg : pos = 18.0 : 1.0 argh = True neg : pos = 17.3 : 1.0 battery = True neg : pos = 16.6 : 1.0

The results are slightly different; this was the case even with Linear SVC. This shows that stopwords really affect the predictions. An intuition to this can be obtained from the fact that given the short length of tweets, people generally use stopwords such as and, while, before, after and so on. Thus removal of stopwords makes a lot of difference to the accuracy.

**Effect of Bigram:**

Bigram uses a combination of two words as a feature. Bigram effectively captures some features in the data that unigram fails to capture. For example, words like ‟not sad‟, ‟not good‟ clearly say that the sentiment is negative. This effect can be clearly seen from the increase in accuracy from 74.56(Unigram) to 76.44 percent which is almost a 2% increase. Following table shows the accuracy obtained at different sizes for the Naïve Bayes algorithm with bigram model.

**Table 6. Accuracy of Naïve Bayes Algorithm (Stopword removal +Bigram)**

|  |  |
| --- | --- |
| **Dataset** | **Accuracy** |
| **10** | 0.544990185582 |
| **50** | 0.593593861527 |
| **100** | 0.591407922912 |
| **500** | 0.654956281228 |
| **1000** | 0.67193076374 |
| **5000** | 0.718214668094 |
| **10000** | 0.730973411849 |
| **15000** | 0.740609386153 |
| **20000** | 0.746431120628 |
| **25000** | 0.75073608137 |
| **30000** | 0.755041042113 |
| **35000** | 0.758453783012 |
| **40000** | 0.762892576731 |
| **45000** | 0.764476266952 |

The most informative features for Naive Bayes with Bigrams as features.

('so', 'sad') = True neg :pos = 55.2 : 1.0 sad. = True neg :pos = 44.2 : 1.0 bummed = True neg : pos = 33.8 : 1.0 horrible = True neg : pos = 32.0 : 1.0

('USERNAME', 'welcome') = True pos :neg = 29.5 : 1.0 ('welcome', 'to') = True pos :neg = 28.1 : 1.0 sad = True neg : pos = 27.5 : 1.0 ('i', 'lost') = True neg :pos = 24.7 : 1.0 died = True neg : pos = 24.3 : 1.0 ('miss', 'him') = True neg :pos = 24.1 : 1.0

**Effect of using Trigram:***.*

Running Naïve Bayes utilizing Trigrams, bigrams and unigrams together gave an delicacy of75.41 percent which is lower than the delicacy attained when Bigrams were exercised as a point. Also this point combination bloats up the point room exponentially and the prosecution becomes extremely tardy. Hence for farther dissection, the trigrams aren't considered as they don't have a notice suitable jolt on the delicacy. Following table shows the delicacy attained at nonidentical sizes for the Naïve Bayes algorithm with Trigram model.

**Table 7. Accuracy of Naïve Bayes Algorithm (Stopword removal+Trigram)**

|  |  |
| --- | --- |
| **Dataset** | **Accuracy** |
| **10** | 0.486995895789 |
| **50** | 0.528484118487 |
| **100** | 0.581571199143 |
| **500** | 0.634346002855 |
| **1000** | 0.654331727338 |
| **5000** | 0.703403818701 |
| **10000** | 0.721002855103 |
| **15000** | 0.731352605282 |
| **20000** | 0.737419700214 |
| **25000** | 0.742148465382 |
| **30000** | 0.74823786581 |
| **35000** | 0.748773197716 |
| **40000** | 0.753234296931 |
| **45000** | 0.754171127766 |

The most informative features for Naive Bayes with Trigrams as features.

('so', 'sad') = True neg :pos = 59.1 : 1.0

('lost', 'my') = True neg :pos = 38.9 : 1.0

('i', 'miss', 'my') = True neg :pos = 36.9 : 1.0

('going', 'to', 'miss') = True neg :pos = 28.5 : 1.0

('miss', 'him') = True neg :pos = 25.4 : 1.0

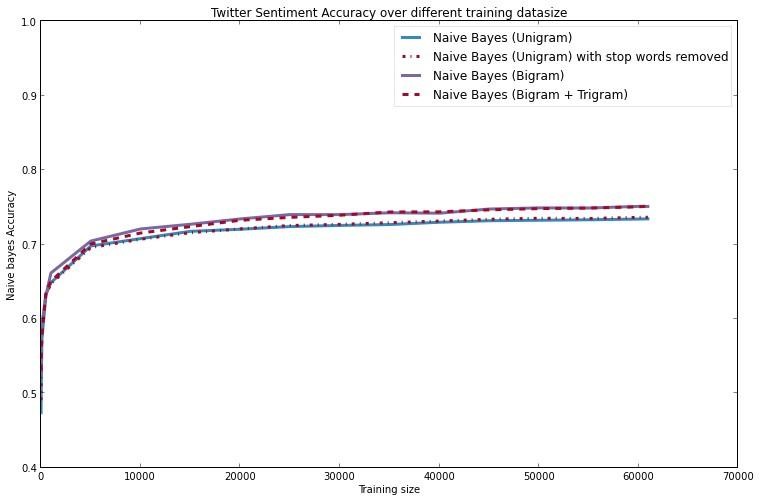
('happy', "mother's", 'day') = True pos :neg = 25.0 : 1.0

("can't", 'sleep') = True neg :pos = 21.5 : 1.0

('sad', 'that') = True neg :pos = 21.5 : 1.0

('miss', 'my') = True neg :pos = 21.4 : 1.0

('i', 'lost') = True neg :pos = 20.9 : 1.0 Following graph shows the summary of the results obtained by using different features and variation in the naïve bayes algorithm.



**Fig.6 Graph Representing Different results obtained for Naïve Bayes Algorithm.**

## Table 8. Accuracy of Naïve Bayes Algorithm

|  |  |
| --- | --- |
| Algorithm | Accuracy |
| Naïve Bayes (unigram) | 74.56 |
| Naïve Bayes (bigram) | 76.44 |
| Naïve Bayes (trigram) | 75.41 |

**C. Support Vector Machine (SVM):**

**Effect using unigram**

Following table shows the accuracy obtained at different sizes for the SVM algorithm with unigram model.

## Table 9. Accuracy of SVM Algorithm (Unigram)

|  |  |
| --- | --- |
| **Dataset** | **Accuracy** |
| **10** | 0.525450571021 |
| **50** | 0.550521948608 |
| **100** | 0.569726980728 |
| **500** | 0.6261375803 |
| **1000** | 0.660421127766 |
| **5000** | 0.726222341185 |
| **10000** | 0.739806388294 |
| **15000** | 0.748973947181 |
| **20000** | 0.75426034975 |
| **25000** | 0.758096895075 |
| **30000** | 0.76130888651 |
| **35000** | 0.762847965739 |
| **40000** | 0.76556923626 |
| **45000** | 0.766862955032 |

**Effect using Bigram**

Following table shows the accuracy obtained at different sizes for the SVM algorithm with Bigram model.

## Table 10. Accuracy of SVM Algorithm (Bigram)

|  |  |
| --- | --- |
| **Dataset** | **Accuracy** |
| **10** | 0.500223054961 |
| **50** | 0.574232690935 |
| **100** | 0.56437366167 |
| **500** | 0.632293897216 |
| **1000** | 0.657989828694 |
| **5000** | 0.725486259814 |
| **10000** | 0.746609564597 |
| **15000** | 0.756468593862 |
| **20000** | 0.761487330478 |
| **25000** | 0.767375981442 |
| **30000** | 0.771011777302 |
| **35000** | 0.77210474661 |
| **40000** | 0.775941291934 |
| **45000** | 0.777324232691 |

## Table 10. Summary for Accuracy of SVM Algorithm

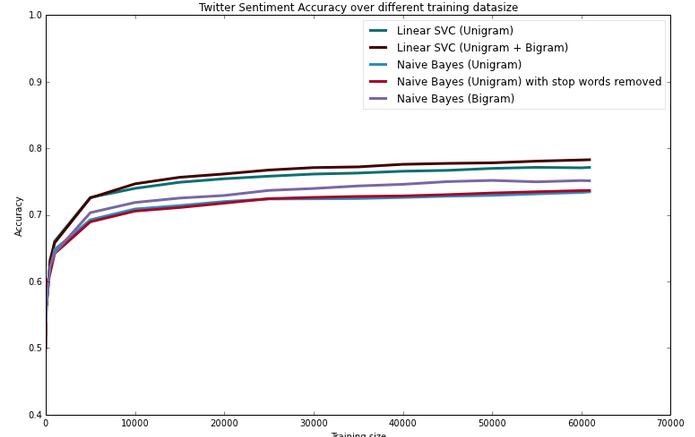
|  |  |
| --- | --- |
| Algorithm | Accuracy |
| SVM with unigram | 76.68 |
| SVM with bigram | 77.73 |

**D. Maximum Entropy**

In Maximum Entropy Classifier, no hypotheticals are taken descrying the relationship betweenfeatures.we attained an delicacy of74.93 percent with unigram model With all the features considered, the effects show off that SVM outperforms NaiveBayes and ultimate entropy as well in all cases. In personal, the point combination of Slang stopwords junking and Bigram gives the ultimate delicacy of77.73 with SVM. ultimate Entropy model gives an delicacy constantly in- between NaiveBayes and SVM. Also it runs iteratively and takes a voluminous quantum of time to run. Hence MaxEnt wasn't exercised for all the point amalgamations. **Table 11. Summary Of Results For Unigram**

|  |  |
| --- | --- |
| Method | Accuracy(Unigram) |
| Baseline | 73.65 |
| Naïve Bayes | 74.56 |
| SVM | 76.68 |
| Maximum Entropy | 74.93 |

As the table shows, when the processing, analysis was done on the bigger dataset, the accuracy scaled upto a great extent. NaiveBayes baseline scaled upto 76.44 and SVM scaled upto 77.73percent.The best result tested thus far, was obtained when SVM was used on a feature set of a combination of Unigram, Bigram with stopwords removal, gave an accuracy of 77.73. MaxEnt also performed well and gave an accuracy of 74.93when stopwords was removed.



**Fig.7 Graph Representing Different results obtained for Naïve Bayes Algorithm And Linear SVC (SVM).**

# CHALLENGES IN SENTIMENT ANALYSIS

Sentiment Analysis is a very challenging task. Following are some of the challenges[13] faced in Sentiment Analysis of Twitter.

**1. Identifying subjective parts of text:**

Subjective parts represent sentiment-bearing content. The same word can be treated as subjective in one case, or an objective in some other. This makes it difficult to identify the subjective portions of text.

For **example**:

1. The language of the Mr.Dennis was very crude.
2. Crude oil is obtained by extraction from the sea beds.

The word „crude‟ is used as an opinion in first example, while it is completely objective inthe second example.

1. **Domain dependence[24]:**

The same sentence or phrase can have different meanings in different domains. For Example, the word „unpredictable‟ is positive in the domain of movies, dramas ,etc, but if the same word is used in the context of a vehicle's steering, then it has a negative opinion.

1. **Sarcasm Detection:**

Sarcastic sentences express negative opinion about a target using positive words in unique way.. ***Example****:*

“Nice perfume. You must shower in it.”

The sentence contains only positive words but actually it expresses a negative sentiment.

1. **Thwarted expressions:**

There are some sentences in which only some part of text determines the overall polarity of the document.

***Example:***

“This Movie should be amazing. It sounds like a great plot, the popular actors , and the supporting cast is talented as well. “

In this case,a simple bag-of-words approaches will term it as positive sentiment, but the ultimate sentiment is negative.

1. **Explicit Negation of sentiment:**

Sentiment can be negated in many ways as opposed to using simple no, not, never, etc. It is difficult to identify such negations .

***Example:***

“It avoids all suspense and predictability found in Hollywood movies.”

Here the words suspense and predictable bear a negative sentiment, the usage of „avoids‟ negatestheir respective sentiments.

1. **Order dependence:**

Discourse Structure analysis is essential for Sentiment Analysis/Opinion Mining.

***Example:***

A is better than B, conveys the exact opposite opinion from, B is better than A.

**7. Entity Recognition:**

There is a need to separate out the text about a specific entity and then analyze sentiment towards it.

***Example*:**

“I hate Microsoft, but I like Linux”.

A simple bag-of-words approach will label it as neutral, however, it carries a specific sentiment for both the entities present in the statement.

1. **Building a classifier for subjective vs. objective tweets.**

Current research work focuses mostly on classifying positive vs. negative correctly. There is need to look at classifying tweets with sentiment vs. no sentiment closely.

1. **Handling comparisons.**

Bag of words model doesn't handle comparisons very well. ***Example****:*

"IIT‟s are better than most of the private colleges", the tweet would be considered positive for both IIT‟s and private colleges using bag of words model because it doesn't take into account the relation towards "better".

1. **Applying sentiment analysis to Facebook messages.** There has been less work on sentiment analysis on Facebook data mainly due to various restrictions by Facebook graph api and security policies in accessing data.
2. **Internationalization [16,17].**

Current Research work focus mainly on English content, but Twitter has many varied users from across.

# APPLICATIONS OF SENTIMENT ANALYSIS

# Passion dissection has numerous operations in colorful Fields. 1. operations that exercise Reviews from Websites Moment Internet has a voluminous collection of reviews and feedbacks on closely everything. This includes product reviews, feedbacks on political effects, commentary about services,etc. thus there is a want for a passion dissection system that can prize sentiments about a personal product or services. It will support us to automate in qualification of feedback or standing for the given away product, item, etc. This would serve the conditions of both the stoners and the merchandisers. 2. Operations as aSub-component Technology A passion predictor system can be helpful in recommender systems as well. The recommender system will not recommend details that allow a lot of inhospitable feedback or lower conditions. In online message, we come across abusive language and other inhospitable rudiments. These can be detected exclusively by relating a largely inhospitable passion and also taking action against it. 3. Operations in Business Intelligence It has been observed that people presently tend to look upon reviews of productions which are accessible online before they buy them. And for multitudinous companies, the online opinion decides the success or failure of their product. thus, passion dissection plays an important portion in companies. Companies also wish to prize passion from the online reviews in order to meliorate their productions and in turn their character and help in customer satisfaction. missions across firmaments Recent inquiries in sociology and other fields like medical, derisions have also been advantaged by passion dissection that show off trends in mortal passions especially on gregarious media. missions In Smart Homes Smart homes are supposed to be the technology of the future. In future exclusive homes would be networked and people would be able to control any portion of the home utilizing a tablet device. recently there has been lot of disquisition going on Internet of goods( IoT). passion dissection would also detect its expressway in IoT. Like for illustration, predicated on the current passion or emotion of the user, the home could revise its air to produce a comforting and peaceful fiefdom. passion dissection can also be exercised in trend prophecy . By shadowing public views, important data descrying deals trends and customer satisfaction can be pulled.

# CONCLUSION

# In this paper, we give a check and relative study of being ways for opinion mining involving engine knowledge and dictionary- predicated approaches, together with trial sphere andcross- lingual styles and some evaluation criteria. Research results show off that engine knowledge styles, analogous as SVM and naive Bayes have the topmost delicacy and can be regarded as the birth knowledge styles, while dictionary- predicated styles are truly operative in some cases, which bear numerous trouble in mortal- labeled document. We also studied the goods of various features on classifier. We can conclude that farther the cleaner data, more accurate effects can be attained. Use of bigram model provides better passion delicacy as assimilated to other models. We can concentrate on the study of combining engine knowledge system into opinion dictionary system in order to meliorate the delicacy of passion type and adaptive capacity to variety of firmaments and nonidentical languages.

# REFERENCES

1. A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326
2. R. Parikh and M. Movassate, “Sentiment Analysis of User- GeneratedTwitter Updates using Various Classi\_cation Techniques",CS224N Final Report, 2009
3. Go, R. Bhayani, L.Huang. “Twitter Sentiment ClassificationUsing Distant Supervision". Stanford University, Technical Paper,2009
4. L. Barbosa, J. Feng. “Robust Sentiment Detection on Twitterfrom Biased and Noisy Data". COLING 2010: Poster Volume,pp. 36-44.
5. Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.
6. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, “Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social

Media,2011 , pp. 30-38

1. Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volumepages 241{249, Beijing, August 2010
2. Po-Wei Liang, Bi-Ru Dai, “Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management,Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1-494673-6068-5, http://doi.ieeecomputersociety.org/10.1109/MDM.2013.
3. Pablo Gamallo, Marcos Garcia, “Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175.
4. Neethu M,S and Rajashree R,” Sentiment Analysis in Twitter using Machine Learning Techniques” 4th ICCCNT 2013,at Tiruchengode, India. IEEE – 31661