Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 41 Issue: 02-2022
DOI 10.17605/OSF.IO/9KZVJ

A SURVEY ON GENERIC VIEW OF SENTIMENT ANALYSIS USING ML

[1]K. DHANA SREE DEVI, [2]CHUNDURU ANILKUMAR and [3]MORSA CHAITANYA

[1]Associate Professor, Department of CSE, CVR College of Engineering, Hyderabad, Telangana, India
[2]Assistant Professor, Department of Information Technology, GMR Institute of Technology, Rajam, AP, India
[3]Assistant Professor, Department of Computer Applications, RVR & JC College of Engineering, Guntur, AP, India

Abstract

Any sentence of a document has two different aspects: the context around which the sentence is constructed and the grammar using which the sentence is build. Any document is weighted and inferred based on the contextual meaning rather than the grammar used for its construction. For in the context we may have deeply hidden opinions, attitudes, diverse emotions, strong sentiments; may be on business products or on social renderings or on raving Human feelings. These opinions may be healthy and positive or unhealthy and negative. A business may expand its sales line by recommending a product to a customer, if it knows the customer opinion on a product prior. Emotions toned behind the series of social posts if analyzed properly may stop several social issues. In politics, to find the views of the voter, towards a specific political group, in quality assurance by finding errors in the products based on past user experiences and many more application scenarios to visit. For a recommender system sentiment analysis has been proven to be a valuable approach. For digging deeper insights sentiment analysis is using Natural language processing (NLP), Text Mining and Machine Learning techniques. There are many sentiment analysis algorithms, but the predominantly used are the machine learning approaches. Though the rule based approach is less accurate in analyzing the sentiments, the rules which it had defined are base to all the Automatic sentiment analysis algorithms. Automatic sentiment analysis algorithms are machine learning algorithms, which uses the word2vec technique, for digging hidden sentiments. This paper includes a generic survey on the sentiment analysis domain. To reach at an understandable level for the young writers this paper presents a review on how to look into any sentiment analysis based machine learning application.

Key words: Sentiment, Sentiment analysis, Text mining, NLP, Machine Learning, word vector, PoS

1. Introduction

Any sentence of a document has two different aspects: the context around which the sentence is constructed and the grammar using which the sentence is build. Any document is weighted and inferred based on the contextual meaning rather than the grammar used for its construction. For in the context we may have deeply hidden opinions, attitudes, diverse emotions, strong sentiments; may be on business products or on social renderings or on raving Human feelings. These opinions may be healthy and positive or unhealthy and negative. A business may expand its sales line by recommending a product to a customer, if it knows the customer opinion on a product prior. Emotions toned behind the series of social posts if analyzed properly may stop several social issues.

As we know dictionary meaning of sentiment is a view or opinion expressed; expressed may be through a write-up document or via a textual post. Sentiment Analysis is mining the contextual meaning of text. As an approach of Natural Language processing [1] sentiment Analysis aims at identifying a written piece of text is positive or negative, customer review on a product is for or against; social posts are these healthy or not. Sentiment Analysis also named as opinion mining is a process of elevating the contextual or subjective opinion of a piece of document or more. These pieces of documents may be sourced from news articles, social tweets, blog posts, business reviews, medical articles, etc.

As we see today's market from any corner is leading using sentiment Analysis. Todays business growth accelerator is sentiment analysis [2]. The increasing social networks and blogs have fueled the business to know more about the customer sentiment on their products. Market research with its Data Analysts is able to understand what customer thinks of its brand; timely monitor brands according to customer sentiments and showcasing successful customer experiences. Business organizations are highly focusing on sentiment analysis to measure how acutely customer opinions are inclined towards their products. Using social media sentiment Analysis [3] organizations are able to track and analyze: popularity of their brand spread by the social communications, a new product if released in the market how it is received and anticipated; the mark of their company's reputation as part of reputation management. To the other corner competitor analysis is also vital for the success. Sentiment analysis is able to capstone market research [4] and is able to provide some handy insights about the competitive products. Voice of the customer is also vital for successful business. Sentiment analysis digs user feedbacks to analyze the customer tone towards a product [5].

There are many social media sentiment analysis tools: picturing some of them here; Hootsuite [6] a social media miner is able to capture 1.3 trillion social media posts, and various demographics that can even attain deeper insights on any product sentiments. Digimind [7] has a core capacity of pulling 850 million web sources and arriving at a comprehensive view of user sentiment on a product. The tool is able to compare the success rate of one brand with that of the competitors. Google Natural Language API extracts sentiments using NLP techniques. Rapidminer uses text mining methods to forecast the user sentiments on a product. The Brandwatch, Mentionlytics, Textrics are some more tools [8] working in the same lines of foreshowing the user sentiments.

Nearly 80% percent of the data from the world is unstructured; with large text messages, emails chats, social tweets. Analyzing such large text quantity is quite difficult. Sentiment analysis can be used to automatically tag such unstructured data [9]. There are many real time situations where sentiment analysis is proven to be useful: emotion recognition and risk prevention, while identifying people being harassed or attacked by analyzing the sentiments of their chats. In politics, to find the views of the voter [10], towards a specific political group, in quality assurance by finding errors in the products based on past user experiences and many more application scenarios to visit. For a recommender system sentiment analysis has been proven to be a valuable approach.

For digging deeper insights sentiment analysis is using Natural language processing (NLP), Text Mining and Machine Learning techniques. There are many sentiment analysis algorithms, but the predominantly used is the Rule based approach. Though the rule based approach is less accurate in analyzing the sentiments, the rules which it had defined are base to all the Automatic sentiment analysis algorithms. To have a deeper view of this special domain sentiment analysis this paper includes the following sections: Section 1: Why sentiment Analysis? How well it works. Section 2: Types of sentiment Analysis. Section 3: Levels of sentiment analysis, Section 4: Sentiment analysis challenges, Section 5: Sentiment analysis algorithms, Section 6: Sentiment analysis: Use case

2. Why sentiment Analysis and how well it works

Sentiment is a view or opinion expressed, expressed may be through a piece of text or via a textual post. Some of these opinions are clear and through meaningful insights. But some of the views are not clearly stated; instead a dig through the views is needed to identify the view is positive or negative. Sentiment analysis comes to the stride at these cases. Whatever may be the domain from which the opinion is conceived from: may be facebook, instagram posts, may be political tweets, may be user reviews on movies, original series, may be customer reviews on products-books, shoes, bags, electronic goods, grocery etc; may be customer feedback on defectiveness; all of these review kinds when harnessed using better methods and tools outputs deeper insights that are productive for the domain growth. Whether business market or social arena sentiment, analysis is a see through lens making sense of polarized opinions and transforming them to beautiful insights useful for decision making.

2.1 Product Analytics

Sentiment Analysis is serving a big deal in Product analytics [11]. Product Analytics is a Business approach for understanding users varied opinions on their branded products, understanding how the users engage and navigate from the product, understanding which feature of the product captures the long retention of the user, analyzing his feedbacks, and how to render for long user retentions. A business forum has showcased that-Companies that use customer analytics comprehensively report outstripping their competition in terms of profit almost twice as often as companies that do not. Market leaders like Netflix, Uber, Zalando are completely into product analytics. For example let us suppose product analytics showed two reviews on iphone as shown:

Review 1: I need an iphone so badly having feature1

Review 2: I just had an iphone, it was so bad with its feature 1

The product analytical tools just shows off these reviews, they just puts these two under one stack and highlight the product talked about is ‘iphone’. But if clearly analyzed the opinions within these two reviews; they are different. Sentiment analysis comes to the rescue of product analysis at these cases. Applying sentiment analysis on these two reviews, review 1 is positive whereas review 2 is negative on feature 1 of the iphone and a look through of feature 1 is needed by the company; otherwise may badly affect iphone market and thus may be useful to their reputation management.

2.2 Public Relations

Public Relations(PR) is another domain [12] where the efforts of Sentiment analysis are intensely reflected. One PR site opens with a message: ‘In developing your public relations program, if you fail to include sentiment analysis, you have denied yourself a vital, dynamic metric to define and promote success’. Public relations community is openly saying that sentiment analysis is an important PR tool. PR is a society managing and rendering services between an organization (business, govt or non govt) and public. Public Relation arenas like entertainment, technology, music, travel, television, food, consumer electronics and more, work at the business to business end. Many businesses use Twitter and facebook as PR tools. The political tweets in twitter are most presumed nightmares of public relations. When the 2020 pandemic crept in, US president has cornered china via his tweets. Reading his tweets roughly seems to be positive; but sentimentalizing them really digs his inner view on china covid. Taking few sample tweets of US president:

Tweet 1: Feb 7

"Late last night, I had a very good talk with President Xi, and we talked about — mostly about the corona virus. They're working really hard, and I think they are doing a very professional job. They're in touch with World — the World — World Organization. CDC also. We're working together. But World Health is working with them. CDC is working with them. I had a great conversation last night with President Xi. It's a tough situation. I think they're doing a very good job.”

Tweet 2: Apr 7

The W.H.O. really blew it. For some reason, funded largely by the United States, yet very China centric. We will be giving that a good look. Fortunately I rejected their advice on keeping our borders open to China early on. Why did they give us such a faulty recommendation?

His tweets from Feb 7 to Apr 7 are with completely varying sentiments. Some days his tweets seem to be positive for china, some days against, some days expressing his sadness. Study using sentiment analysis showed upcoming US elections are the sole impact for these fluctuations. Figure shows

2.3 Customer Service

Customer Service is another area where sentiment analysis is predominantly applied [13]. Most of the business targets are customers and their retention. Customer service provides various kinds of services to the customer before and after he buys the product. Bad customer services can account to the fall of organization. Forbes reports 2016, revealed that organizations are losing about $15 billion a year because of poor customer service; this is because customer emotions are not properly understood may be by the agent or organization; their hidden sentiments towards the products are not clearly taken over by the agent or organization. Having lost such huge business companies today are prioritizing customer service more than product price; because customer service if analyzed properly may lead to retention of customers to the product. Today customer service is using Sentiment analysis to create healthy emotional experience that drives customer loyalty towards the product. Big Deals like Amazon, Netflix, Flip cart are using the strategic decisions of sentiment analysis to retain their customers. Let us for example take the some of the messages exchanged during customer service of a company before and after sentiment analysis is actually applied.

Conversation 1:

Jessica: Thanks for holding. Logimax support. This is Jessica.

Customer: Yeah, your company promised me delivery on Tuesday. It's now Friday and it's still not here. Do you have any idea what kind of problems you're causing me?

Jessica: Sir, please, Calm down!

Customer: Calm down?!? Calm down??? Don't tell me to calm down!!! Thanks for the advice, but if your company would have done what they promised, we wouldn't even be having this conversation!

Conversation 1 shows the customers irony towards Logimax just because Jessica used the word calm down. From Jessica's side using the word may be normal, but we don't know how the customer emotions change for such words and may sometime negatively polarize his emotions. The subject of the conversation is changed within no time. The outcome may be Jessica will lose her customer. Applying sentiment analysis companies came to know that, telling a customer how to act is never a good idea. The application analyzed capturing the customers concerns and emotions is the first target for any company. Lets us see a conversation scenario after the company is trained on sentiment analytics.

Conversation 2:

Customer: Yeah, your company promised me delivery on Tuesday. It's now Friday and it's still not here. Do you have any idea what kind of problems you're causing me?

Jessica: My apologies. That's got to be very frustrating. If need be, we can overnight your order so it's there by Monday. Now please tell me exactly what happened and I will get to work on fixing this.

Customer: Yes OK. Thanks….

What might have happened now; the customer in conversation 2 is completely biased towards Jessica; this is where sentiment analyses is applied. Jessica was trained to use positively polarized words like apologies, please, fixing this. Jessica was successful in retaining her customer. Millions of such conversations happen by the end of a day; and sentiment analysis always digging new insights on tackling such sudden customer emotions

3. Types of sentiment Analysis

In the context of business the customer emotions vary as product or brand vary. Sentiments may have polarized variations or emotional variations or varied intentions.

Organization has to know the varied sentiments of the targets. In this section we focus on most popular types of sentiment analysis [14].

3.1 Standard sentiment analysis

This approach uses the direct meaning of the sentiment and categorizes it to be positive or negative. Most popularly used. For example consider feeds on Government during the pandemic 2020:

1. I like the way govt is taking different steps during this pandemic and bringing the world together → Positive
2. I still need to analyze govt steps are useful to the nation or not →Neutral
3. govt acts during the Pandemic are so confusing → Negative

A direct understanding of the meaning of these 3 opinions categorizes them to be Positive, Neutral and Negative. Here categorizing the feeds to Positive and Negative showcases the depth of the emotion and here it’s not that deep.

3.2 Fine grained sentiment analysis

If the polarity is chosen as precision metric then the intensity of the polarity can be expanded to categories holding higher precisions like: very positive, positive, neutral, very negative, negative. To determine the sentiment of the text having these finer granules; fine grained sentiment analysis is used [15]. They showcase the deeper intensity of the sentiment. For example consider the other set of feeds on the govt during the pandemic 2020:

1. The older govt was much needy-----Negative on the current govt
2. Awful govt decisions. Tax is not exempted. It’s painful for employees like us during this crisis. I will not vote for this govt again……..Very negative on the current govt
3. I don’t think there is anything I really dislike about the govt……..Neutral

Feed 1 showcases negativity. Feed 2 on the other hand showcases the depth of the opinion and is completely harmful to the govt as here the govt is losing the voter. Such types of sentiments are to be carefully analyzed for the success of any organization or company.

3.3 Emotion Detection

This model detects the hidden emotions underneath a text. Emotions of happiness, sadness or anger[16]. Emotions inevitably affect human decision making. Humans tend to repeat the actions that make them happy and tend to avoid that which making them sad or angry. For example consider the set of feeds as shown:

F1:‘govt free internet service makes my day a lot easier :)’….. Happiness
F2:‘govt free internet customer service is a nightmare! Total dump and totally useless!!!’…. Anger

In feed 1 usage of the words ‘make my day a lot easier’ identifies the emotion underneath is Happiness. In feed 2 usage of the words ‘nightmare’, ‘totally dump’ and ‘useless’ identifies the emotion underneath is Anger.

F3:‘This offer from the govt during this pandemic is not friendly’ …negative sentiment/emotion of sadness
F4:‘I hate this pandemic offer. Its worst’…Negative sentiment/ emotion of anger

Both of the feeds 3, 4 are categorized to negative sentiments. But both are emotionally different. Feeds of type 4, needs an executable action, dragging the voters to friendly emotions by offering what they like.

3.4 Aspect Based Sentiment Analysis

Aspect is a feature of an object or an entity under discussion. This sentiment analysis analyzes and understands the feature of an entity focused in the customer opinion. Aspect based sentiment analysis tries to associate the sentiments to various aspects of a product or a service [17], thereby harnessing more detailed and accurate hidden relations. Product reviews many a times includes with different opinions on different features or characteristics of a product like price, structure, Interface design etc. For example consider the feeds as shown:

F5:‘Web extensions of Todoist!! Most wonderful, I like them. A great and easy to use tool to collaborate with far off experts’
F6:‘Web extensions of Todoist may be awesome!!! But I bet cannot compete Evernotes services’

Both of feed 5 and 6 are talking about entities which are two productivity tools for work from home: Todoist, Evernotes.

Feed 5 harness 95% positive opinion with words: most, wonderful, like them, great easy, collaborate. Feed 5 is an aspect based opinion where the aspect is web extension. The discussion is 50% positive to the feature alone. Sentiment analysis just focuses on the positivity of the feed. But Aspect based sentiment analysis analyzed the context is on web extension feature and user is more inclined to it. The decision making can be, providing more of such extensions so as to retain the customer for this product and enhance sales of the product.

Feed 6 on the other hand is a negative sentiment. More than the sentiments if we look the aspects: feed 6 is comparing two aspects which are services provided by two entities: Todoist and Evernotes. The sentiment is more positive to the Evernote aspect. The decision making can be incorporating similar services of Evernotes into Todoist; and control customers shift to other services.

3.5 Intent Detection

Many of the feeds and reviews if studied deeply incorporate some kind of action may be: to be done or already existing or giving raise to. Intent detection type of sentiment analysis tries to find the hidden action behind an opinion; something the user want to do or something the user expects. Knowing the customer intentions within his feeds and providing services before he actually request, definitely leads to the company’s growth. For example consider the feeds as shown:

F7:‘Hello, I’m an event manager, working with huge amount of raw files. May I know what kind of storage do you offer? Is it being for lifetime? I would like to know more’

F8:‘My day started with frustration. Gmail keeps closing when I log in. Can you help?

In feed 7 the customer is communicating with the cloud service. The sentiment within the feed is neutral showing nothing. The intent detection identifies the customers buying intent of the storage. Now the intent showcases the customer wish to buy their storage provided their services are better.

In feed 8 the hidden intension is request for assistance. If the company immediately identifies and provides the needed service customer retention works.

3.6 Tools servicing kinds of sentiment analysis

The kinds of sentiment analysis discussed above are really great deals to any company to think for their growth. Adopting a sentiment analysis tool benefits the organization in taking the timely decisions.

Dialogflow an intense based sentiment analysis tool uses Natural language API to perform the intent analysis. ParallelDots is another intent analysis API working with advanced services like Bot and spam removal. The API is using LSTM to classify the intent in the text.

NetOwl is an aspect based sentiment analysis tool. Its key feature is capturing multiple conflicting sentiments linking to an entity. AYLIEN is another aspect based sentiment analysis tool with a key feature of diving deep into customers opinions.

There are many impressive emotion detection sentiment analysis tools available. IBMs WATSON tone analyzer is able to identify the emotional tone within a written text. It used SVM as an internal emotion classifier. Qemotion is a text to emotion API, with a key feature of generating an emotional index for a text submitted.

4. Levels of sentiment analysis

Depending on the purpose and the problem need sentiment analysis algorithms can be applied to various levels [18] of a text. These various levels are:

- Document level
- Sentence level
- Entity level

4.1 Document level

The size of a document is complex with many incomprehensible facts and opinions. Document level sentiment analysis finds the overall opinion of the document. At this level each document expresses opinion on a single entity. Research showed many machine learning methods, both supervised and unsupervised applied at document level. Internet

Blogs, online reviews all these are unstructured larger documents where from sentiments are to be extracted.

There are many API available for document level sentiment analysis, using which extensive quantity of text can be processed with minimum delay and high accuracy. The Microsoft Text Analytics API provides powerful web services for extracting sentiments within larger unstructured documents. This API has features like: sentiment analysis, language detection, entity linking, and extracting key phrases. The API uses cloud based machine learning algorithms backend.

The Meaning cloud sentiment analysis API is another powerful sentiment extractor. The API expresses the sentiment as positive or neutral or negative, by analyzing the local polarities within text and identifying relations between these polarities. This API can accomplish tasks such as detecting irony in sentences, differentiating opinions and facts, identifying ambiguities.

The Human Like Sentiment Analysis for Hotel Reviews API allows to extraxt the sentiments hidden in user reviews, for Hotels particularly. Besides extracting sentiments it also projects useful score on the user reviews which may be useful for throwing recommendations. This API is texted to show 95% accuracy. Besides extracting sentiments this API is proved to be helpful in semantic classification and comparison of two hotels and also in recommending good lot of similar hotels.

The Intellexer API includes a suite of tools using which the users can integrate Natural language processing and many other text mining capabilities into their ongoing application. This API uses efficient hybrid techniques that combine linguistic analysis and statistical analysis with various semantic rules. This API is powerful and efficient in analyzing varied expressions, opinions both contextual and non contextual.

The Alchemy Text API an IBMs release and offers better solutions for sentiment analysis. The other powerful feature of the API is analyzing the emotional tone hidden in the content. The API process huge text documents and segregates emotive phrases within them and assigns a polarity score between -1 to 1 and based the score classifies document sentiment as positive or negative. Apart from sentiment analysis this API also includes other capabilities such as concept extraction, keyword extraction, entity extractions.

Today with the advent of Machine learning and deep learning various sentiment analysis models at document level have crept in. Naïve bayes and support vector machines have proved accurate for unstructured review sentiment classification. But failed to extracts relational sentiments in emotions and failed at various levels of ambiguities. The deep neural networks are successful at these situations.

4.2 Sentence level

Finer granule sentiments can be analyzed at sentence level. Documents are divided into sentences. At sentence level we can find the polarity of a sentence whether positive or negative. Many of times, document level sentiment analysis failed to extract hidden sentiments within the larger documents. Sentence level algorithms understands document sentence by sentence, there by the contextual sentiments are better understood. There are many API available to extract sentiments at sentence level. Many of Document level API are also used to extract opinions at sentence level.

The Twinword sentiment analysis API is used to identify the tone of a sentence. It includes other features like finding out the mood hidden behind a written sentence, text classification, finding word associations, and finding sentence similarities.

The Bitext sentiment analysis API uses deep linguistic analysis to arrive at better sentiments. The API uses robust parsing technique to carryout sentiment analysis at both sentence level and at phrase level. In doing so the API showed off better accuracy. Beside sentiment analysis the API is also highlighted for detecting the topic of the sentiment, measuring the intensity of the sentiment, and performing entity extraction.

4.3 Entity level

Using document or sentence level sentiment analysis doesn't always extract true opinions. Applying at finer granules may analyze useful sentiments. Entity level sentiment analysis extract accurate and detailed insights by digging deeper than sentence level. The entity level sentiment analysis first identifies the entities within a document and then analyzes the sentiments between various entities. Remember entities are different from words within a document.

The cloud Natural language API combines both entity analysis and sentiment analysis and finds the varied sentiment between various entities. For each entity a sentiment score is determined and based on this score the polarity is labeled as positive or negative.

5. Sentiment analysis challenges

Diving deep into the domain of sentiment analysis, we can find out many pitfalls faced by the analysis. There are lot of serious problems which badly affect the accuracy. Even today many of the sentiment analytical tools are suffering from completely showcasing the gist of a document. In this section we present some challenges faced by sentiment analysis:

1. Negation Detection
2. Sarcasm Identification
3. Word ambiguity
4. Multi-polarity

5.1 Negation detection

Negation in linguistics means which reverses the text polarity. Some Negation words which are encountered by sentiment analysis tools are: Not, Never, No, etc. The contextual polarity of the word is completely changed with these Negation words. But the presence of one single negation word in the text doesn't mean the complete text is inverted and it even doesn't mean that the absence of negation word the text is positive. This is a challenging issue faced by sentiment analysis. For example consider the feeds:

F9: ‘ Such a behavior, it will be my first and last meet’
F10: ‘This is not good from his side, he never thinks narrow. We will wait for the next meet’

Feed 9 uses no negative words, but still from common sense the feed is totally negative. Feeds of this type are at critical analysis for sentiment analytics. Feed 10 on the other hand uses one negation word: Not. The usage of this one ‘Not’ shouldn’t identify the feed is Negative. Clearly feed 10, says the past behavior of the person is good and that momentary behavior is bad and maybe he changes later. Handling such text is a major challenge faced by sentiment analysis.

Latest research addressed this drawback using Recurrent Neural Network Models and LSTMs and outperformed other models with better accuracy.

5.2 Sarcasm Identification

Sarcasm is completely negative as of dictionary meaning. Sarcasm is where people express negative opinions using positive words. Sentiment analysis models are greatly cheated by text of sarcasm [19]. Unless the tool completely understands the context of the situation it cannot correctly dig out the polarized sentiment. For example consider the feeds:

F11: ‘We drove very fast at Jet speed’
F12: ‘ We drove very fast at the speed of Tortoise’

Feed 11 is non sarcastic. Feed 12 is sarcastic. Sarcastic comparison to the speed of tortoise negatively polarized feed 12. Text like this cheats the model accuracy.

Automatic sarcasm detection using Deep Neural Networks is gaining popularity. Both of convolution neural network and LSTM architectures are used to reach out higher accuracy.

5.3 Word Ambiguity

Word ambiguity is another pitfall degrading the performance of Sentiment analysis. Word ambiguity is where: two words mean the same or same word changes the meaning of the text. The second case has to be carefully dealt with. When the words are ambiguously defined in the text, it is completely difficult to analyze the polarized opinion. For example consider the feeds:

F13: ‘The story is unpredictable’

F14: ‘Life’s story is unpredictable’

Both of the feeds are talking about stories. But feed 13 is talking about a movie story or a novel. The word unpredictable in feed 13 shows the person is happy with the story and weights for positive opinion of the feed. Feed 14 is talking about the turns in the life. Usage of unpredictability in the life generally shows a frustrated mind and thus here the same word weights for a negative opinion.

Lexicon based sentiment analysis models are well trained to resolve word ambiguities.

5.4 Multi Polarity

Multi polarity is very commonly seen in large customer reviews [20]. A case where a text or document shows more than one polarity is identified to be exhibiting multi polarity. Text can be containing many subjects. A person can praise one subject and comment a different subject occurring in the same document. Many a time's sentiment analysis failed at these situations. For example consider the feed:

F15: 'the story line of the movie is quite interesting, but the direction was too clumsy'

Feed 15 talks about a movie review on two subjects: the story line, direction. The opinion on the first subject is positive whereas on the second, its negative. Here the feed 15 is exhibiting multi polarity.

6. Sentiment analysis algorithms

Multiple approaches model sentiment analysis; two best approaches are:

1. Rule based analysis
2. Automatic sentiment analysis.

Both of the approaches use various natural language processing derivatives to dig sentiments of documents. This section presents the details stages in these approaches.

6.1 Rule based Approach

The rule based approach use rules [21] [22] framed from Natural Language processing. The techniques used in are:

- Tokenization
- Stemming/lemmatization
- Parts of speech tagging
- Parsing

The rule based approach works as follows:

Initially before the approach starts we have to define two word lists: the negative word list and the positive word list. The negative word list includes the words of negative meaning like: bad, ugly, sad, worst etc. the positive word list includes words of positive meaning like: good, happy, excellent. These word lists are fed to the rule based model at first. When the user inputs his document for analyzing the sentiments the model applies the techniques mentioned above.

6.1.1 Tokenization

Tokenization is the process of breaking the documents into words or keywords or phrases and these atomic elements are called tokens [23]. While tokenizing a document special characters like: !, ?, &,$ etc are discarded. Tokenization plays a vital role in lexical analysis. Tokenization can be at sentence level also where the document is broken to sentences.

Code 6.1.1.1: Sentence level

```
from nltk.tokenize import sent_tokenize
text1 = " My day started with frustration. Gmail keeps closing when I log in. Can
you help "
sent_tokenize(text1)
```

Output 6.1.1.1:

```
[‘My day started with frustration.’,
‘Gmail keeps closing when I log in.’,
 ‘Can you help’]
```

Code 6.1.1.2: Word level

```
from nltk.tokenize import word_tokenize
text2 = "We drove very fast at Jet speed!. Speeding my drive saved time, I am best
driver, good management"
word_tokenize(text2)
```

Output 6.1.1.2:

```
['We', 'drove', 'very', 'fast', 'at', 'Jet', 'speedSpeeding', 'my', 'drive', 'saved', 'time',
'I', 'am', 'best', 'drivergood', 'management']
```

Code 6.1.1.1 uses sentence level tokenization on the document text1 and code 6.1.1.2 uses word level tokenization on the document text2.

6.1.2 Stemming/Lemmatization

Both of stemming and lemmatization are the techniques for transforming tokens to root words. They work by converting the word to its base form by removing the affixes. This conversion is needed because tokenization after breaking into tokens doesn’t check for the affirmative words which are almost the same. Replacing all the affirmatives to a single map decreases the data size.

The purpose of stemming and lemmatization is same but lemmatization is different from stemming. Both of them replace affirmative words to base word. The base words produced by stemming don’t preserve the contextual meaning. When we refer to sentiment analysis contextual meaning of a word is vital. On the other hand lemmatization preserves the contextual meaning. For example in code 2 of 6.1.1, output2 tokens include affirmative words: ‘drove’, ‘drive’ and ‘speed’, ‘speeding’; these pairs convey the same meaning. Stemming and Lemmatization replaces two words with single base word so as to reduce the corpus size. Stemming cannot assure a language word but lemmatization always results in a language word which preserves its context. These techniques are called as Text normalization techniques. For example consider the code shown:

Code 6.1.2.1: Stemming

```
from nltk.stem import PorterStemmer
stemmer = PorterStemmer()
stemmer_words=[stemmer.stem(token) for token in tokens]         # tokens from output2
print(stemmer_words)
```

Output 6.1.2.1:

```
['We', 'drove', 'veri', 'fast', 'at', 'jet', 'speedspeed', 'my', 'drive', 'save', 'time', 'I', 'am',
'best', 'drivergood', 'manag']
```

Code 6.1.2.2: Lemmatization

```
lmtzr = WordNetLemmatizer()
lmtzr_words = [lmtzr.lemmatize(token) for token in tokens]
print(lmtzr_words)
```

Output 6.1.2.2:

```
['We', 'drove', 'very', 'fast', 'at', 'Jet', 'speedSpeeding', 'my', 'drive', 'saved', 'time',
'I', 'am', 'best', 'drivergood', 'management']
```

Analyzing outputs 6.1.2.1 & 6.1.2.2 of stemming and lemmatization: stemming changed the word management to the root word ‘manag’ which doesn’t show any contextual meaning; on the other hand output 6.1.2.2 shows lemmatization preserved the word ’management’ inorder to retain the contextual meaning.

6.1.3 Parts of speech tagging

Parts of speech tagging (POS) is a technique where token pos is identified and tagged to it. Deeper sentiments can be analyzed only if known how a word is used in the sentence. PoS in general shows how a word is used in the sentence. Various PoS are: Noun, Verb, adverb, adjective etc. PoS is essential in sentiment analysis. For example consider the feeds and POS codes:

F16: ‘I like what ravi does’

F17: ’Do it like you know’

Code 6.1.3.1: POS tagging F16

```
F16=re.sub(r'\d|\.|\,|\?|\!','',F16)          # Removing numbers, fullstop and comma
```

```
tokens1 = nltk.word_tokenize(F16)
tokens1
print(nltk.pos_tag(tokens1))
```

Output 6.1.3.1:

```
[('I', 'PRP'), ('like', 'VBP'), ('what', 'WP'), ('ravi', 'NN'), ('does', 'VBZ')]
```

Code 6.1.3.2: POS tagging F17

```
F17=re.sub(r'\d|\.|\,|\?|\!','',F17)          # Removing numbers, fullstop and comma
tokens2 = nltk.word_tokenize(F17)
tokens2
print(nltk.pos_tag(tokens2))
```

Output 6.1.3.2:

```
  [('Do', 'VB'), ('it', 'PRP'), ('like', 'IN'), ('you', 'PRP'), ('know', 'VBP')]
```

Comparing output 6.1.3.1 and 6.1.3.2, in output 6.1.3.1 the word ‘like’ is tagged with verb in singular present tense (VBP). A verb may identify the sentiment either positive or negative, based on the context in which it is occurring. Here it weights positive for F16. On the other hand the same word ‘like’ is tagged as an interjection in output 6.1.3.2. Here we have to understand that the tool is tagging a word based on the context and not by its dictionary meaning. Lets us consider one more feed for showcasing why PoS is needed for sentiment analysis:

F18: ‘terrible pitch and awful hitting led to the loss'

Applying PoS tagging on F18, the output is:

Output 6.1.3.3:

```
[('terrible', 'JJ'), ('pitch', 'NN'), ('and', 'CC'), ('awful', 'JJ'), ('hitting', 'NN'), ('led',
'VBD'), ('to', 'TO'), ('the', 'DT'), ('loss', 'NN')]
```

Named entities are general english nouns or pronouns: book, jay, pitch etc. Adverbs or adjectives describe the actions on these nouns or pronouns. In sentiment analysis any of such action from a noun or a pronoun hides some opinion or sentiment. Sentiment analysis system gains some sentiment clues by identifying these adjective-noun pairs. In F18 one such pair is: ('terrible', 'JJ'), ('pitch', 'NN'). Comparing this to Negative list of words which are fed initially to the system, usage of terrible identifies a negative sentiment is within F18. In turn as a human reader we can say F18 has negative sentiment or opinion. Figure 6.1.3.1 shows parts of speech tagging on feed 18.

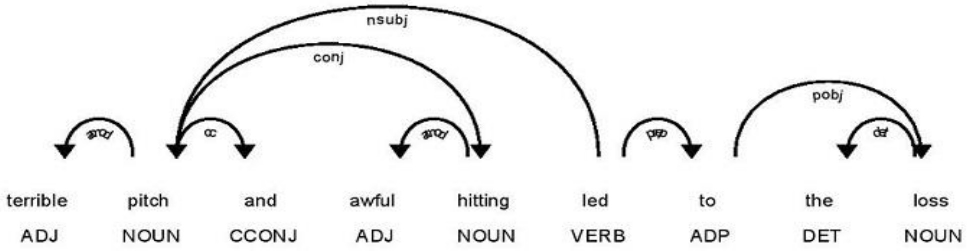

Figure 6.1.3.1: PoS tagging and dependency parsing

6.1.4 Parsing

In general parsing is syntactical analysis. Syntactical relations many a times produce semantic relations. Deep parsing outputs grammatical dependency relations between words. In figure 6.1.3.1 the dependency relations are shown using a directional arrow. Dependency relations, when analyzed show hidden sentiments.

6.1.5 Working of Rule based algorithm

After building the positive and negative list of words the algorithm is trained with these two lists. Once the document to analyze the sentiment is given input to the algorithm the algorithm applies all the above discussed rules and compares the extracted words, relations, features with the available two lists. The algorithm calculates the frequency of positive and negative words within the document. Based on the higher frequency of positive or negative words, the algorithm polarizes the document sentiment as positive or negative. For example if there are more positive words and features the algorithm polarizes the document as positive.

One of the major pitfall of rule based is it uses word matching technique to identify the frequency of positive and positive words in the document. This may not be an accurate approach for finding the actual document sentiment. For example consider the words: happy and not sad. Both of these present a positive sentiment and hence the frequency count has to be 2. But when word matching is applied the word not sad is not counted as positive. These ambiguous words badly affect the rule based approach.

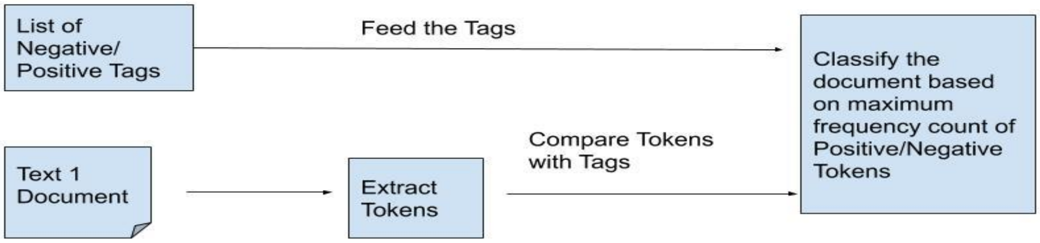
List of Negative/ Positive Tags
Feed the Tags
Classify the document based on maximum frequency count of Positive/Negative Tokens
Text 1 Document
Extract Tokens
Compare Tokens with Tags

Figure 6.1.5.1: working of rule based algorithm

Rule based algorithm is proved to be accurate for finer grained and aspect based (discussed in 3.2.1 & 3.4) sentiment analysis approaches, where word matching is the only technique applied by the method. But it produced low precisions in case of emotion and intent based sentiment analysis (discussed in 3.3 & 3. 5). Word matching is not a good approach incase of ambiguous words and varied intents. It also suffered from accuracy failures while sentimentizing documents with sarcasm, word ambiguities, and multiple polarity words. Figure 6.1.5.1 sows the working of rule based algorithm.

6.2 Automatic sentiment analysis algorithms

Automatic sentiment analysis algorithms use machine learning and deep learning techniques like classifications [24]. Today they are into big deals as they truly dig deeper into the documents to find hidden sentiments and succeeded in the areas where ruled based failed. These algorithms used three popular machine learning stages:

- Data preprocessing stage
- The training phase
- Testing phase

Data preprocessing stage is very import for accurate model building. The data preprocessing stage uses the rule based steps to clean and transform the document to tokens, where stop words are removed. Automatic sentiment analysis algorithms from machine learning domain don’t use word similarity techniques to identify the sentiments. We have seen in rule based algorithm how the word matching technique fails. One of the efficient techniques used by these algorithms is transforming words to feature vectors, which are numeric. A similarity metric is used to identify similar vectors in other wise: similar words. This process of conversion of words to vectors is called feature extraction. Many of the machine learning models uses three basic techniques which convert word to vectors:

- Bag of words(BoW)
- Word2vec
- Term frequency inverse document frequency (Tf-Idf)

The training phase is a learning stage by the algorithm where it learns tagging the positive and negative words. The training phase also includes learning with all the feature vectors, the positive and the negative feature vectors of the sentiments. In the testing phase also called as prediction phase the model which is build in the training phase is tested against unknown document. The feature vector for this new document is generated and compared with the feature vectors of the trained documents and is tagged with most similar document label. The accuracy of the model depends on the number of correct predictions. Figure 6.2.1 shows the working of a machine learning algorithm.

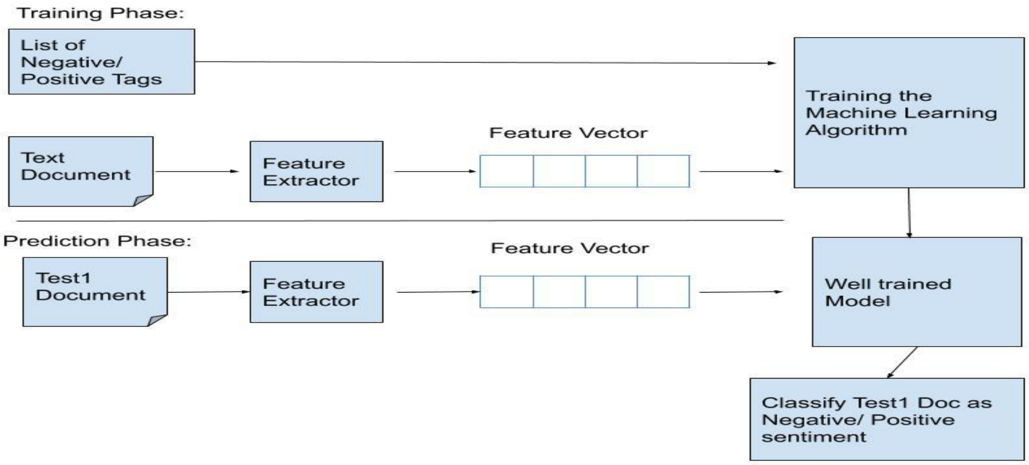

Figure 6.2.1: Working of Machine learning algorithm

Many algorithms contribute to sentiment analysis and some of the important categories are:

- Regression
- Naïve Bayes
- Support vector Machines
- Neural Network variants like: RNNs and LSTMs.

Here in this section we focus on how Recurrent Neural Networks are applied in sentiment analysis.

6.3 Recurrent Neural Networks in Sentiment analysis

With the advent of neural networks [25], Natural language processing is made easier. Variants of neural networks like recurrent neural networks, Long short term memory units [26] [27], addressed many unresolved problems of sentiment analysis. The input and outputs in a tradition neural network are independent of each other. Recurrent neural networks (RNN) are a class of neural networks where the outputs from the previous layer are recurrently fed as input to the current layer (loop). This type of connection acts as a memory to the current layer from the previous layer. When related to text processing the previous layers has some words; when the current layer is building a sentence the previous words are needed to predict the next word of the sentence; hence a connection in the design. In RNNs the hidden layers are the memory units, and when larger documents are processed these hidden layers increase in numbers thus enhancing a large memory unit. This feature of remembering the previous inputs is called Long Short term Memory (LSTM).

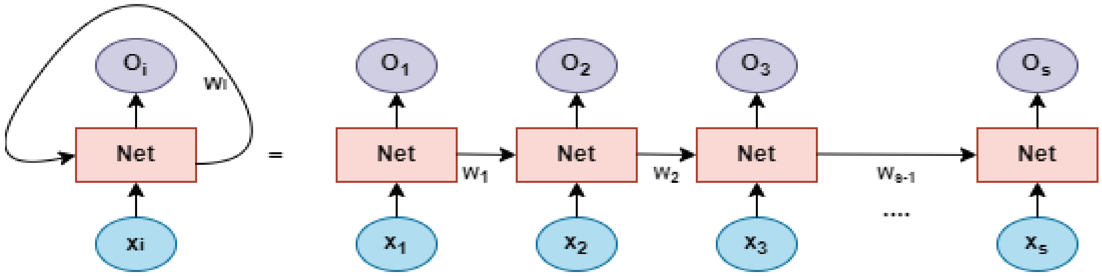

Figure 6.3.1: Unfolding recurrent neural network

The RNNs are typically designed to recognize the sequential characteristics of data and are most popular in text mining, NLP and speech recognition. They are majorly used in sentence automation (auto type) where prediction of next occurrence word is implicitly identified from the previous words. Many of these supervised algorithms uses pre trained libraries for sentiment analysis. Figure 6.3.1 shows unfolding of recurrent neural net layers.

7. Sentiment analysis-Use Case

In the previous sections we have seen rule based and machine learning supported automatic sentiment analysis algorithms. It’s worth nothing if machine learning algorithms are explained without a piece of code. This section presents a real time use case on analyzing the sentiments of movie reviews. For this use case we will use two datasets the Cornell university Movie Review dataset and reviews for an Amazon product. Each review is tagged with a label indicating whether it is a positive (pos) review or negative (neg) review. The Amazon product reviews are collected for a Zoya Mic, where a scraper is used to collect the same. The second dataset has no feature indicating the sentiment labels. We used Vader sentiment analyzer tool append the label feature to this dataset.

Some of the documents are included with NULL values. We removed them. The classification labels are equally distributed showing there is no class imbalance. A train/test split of 70% is applied.

7.1 Evaluate performance

The performance of two models is studied experimentally. We used Naïve Bayes and LSTM classifiers. The classification report of both is shown below

Naïve Bayes-Dataset1

```
precision   recall  f1-score  support

neg            0.84        0.84     0.84        308
pos            0.85        0.85     0.85        332

accuracy          0.85        640
macro avg          0.85        0.85     0.85       640
weighted avg    0.85        0.85     0.85       640
```

LSTM –Dataset1

```
               precision   recall  f1-score  support

neg            0.89        0.86    0.88       310
pos             0.87        0.85    0.87        330

accuracy      0.88        660
macro avg      0.87        0.87      0.87      640
weighted avg  0.87        0.87      0.87       640
```

Naïve Bayes-Dataset2

```
   precision    recall   f1-score   support

      neg           0.74          0.80      0.82        410
      pos           0.81          0.81      0.83        520

 accuracy           0.87          930
 macro avg          0.87          0.85       0.85       930
 weighted avg       0.87          0.85       0.85       930
```

LSTM-Dataset2

```
              precision     recall        f1-score      support

neg           0.90          0.88          0.88          410
pos           0.86          0.89          0.88          520

accuracy        0.88        930
macro avg       0.87        0.87          0.87          930
weighted avg    0.87        0.87          0.87          930
```

From the classification report the LSTM model performed well on both the datasets.

8. Conclusion

Business organizations are expanding their market levels using sentiment analysis. Sentiment analysis is a Natural language processing technique of identifying whether the opinion of a person is positive or negative; also called as opinion mining. The increasing social networks and blogs have fueled the business to know more about the customer sentiment on their products. Market research with its Data Analysts is able to understand what customer thinks of its brand; timely monitor brands according to customer sentiments and showcasing successful customer experiences. Business organizations are highly focusing on sentiment analysis to measure how acutely customer opinions are inclined towards their products. Digging deeper emotions of social network users is today's market focus, and is achieving using sentiment analysis. There are many real time situations where sentiment analysis is proven to be useful: emotion recognition and risk prevention, while identifying people being harassed or attacked by analyzing the sentiments of their chats. In politics, to find the views of the voter, towards a specific political group, in quality assurance by finding errors in the products based on past user experiences and many more application scenarios to visit. For a recommender system sentiment analysis has been proven to be a valuable approach. All these are done using various sentiment analysis approaches like traditional rule based method and using machine learning methods. Machine learning sentiment analysis approaches have proved to extract useful sentimental information from varied documents.

References:

[1]. J.Yi, T.Nasukava, W.Niblack,"Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques", third IEEE International conference on Data Mining, Nov,2003.

[2]. H.Chen,RHL Chiang, " Business intelligence and analytics: from big data to big impact", Management Information Systems, Vol.36,No.4, Dec 2012.

[3]. Y Yu,W Duan,Q Cao," The impact of social and conventional media on firm equity value: A sentiment analysis approach", Decision support systems, Elsevier,Vol.55,No.4,Nov,2013.

[4]. Barrier Gunter, Nelya Koteyko, D Atanasova, " sentiment analysis: A market relevant and reliable measure of public feeling, SAGE journals, Vol. 56,No.2, Mar,2014.

[5]. CP Li,LH Guo,N Lin,” Value Mining of product reviews based on sentiment analysis”, Journal of applied Mechanics and Materials, Vol.713,2015.

[6]. W He,H Wu,G Yan,” A novel social media competitive analytics framework with sentiment benchmarks”, Information and Management, Elsevier, Vol.52, No.7, Nov,2015.

[7]. Mouna EI MArrakchi, Hicham Bensaid, M Bellaf,” Scoring reputation in online social networks, IEEE Xplore, Dec,2015.

[8]. R A S C Jayasanka, M D T Madhushani, E R Marcus, I A A U Aberathne, “ Sentiment analysis for social media”, researchgate, 2013.

[9]. V Subramaniyaswamy, V Vijayakumar, “ Unstructured data analysis on big data using map reduce”, Procedia Computer Science,Vol.50,2015.

[10]. J Ramteke, S Shah, D Godhia,” Election result prediction using Twitter sentiemnt analysis”, ICICT, Aug,2016.

[11]. W Wei, JA Gulla,” Sentiment learning on product reviews via sentiment ontology tree”, Proceedings of the 48th annual meeting of the association for computational linguistics, July, 2010.

[12]. M Bautin, L Vijayarenu, S Skiena,” International sentiment analysis for news and blogs”, ICWSM, 2008.

[13]. A Bagheri, M Saraee, F De Jong,” Care more about customers: Unsupervised domain independent aspect detection for sentiment analysis of customer reviews”, Knowledge Based Systems,Elsevier, Vol. 52, Nov, 2013.

[14]. Neha Raghuvanshi, J M patil,” A brief review on sentiment analysis”, International conference on Electrical, Electronics and Optimization Techniques, Mar, 2016.

[15]. M Van de Kauter, D Breesch, V Hoste,” Fine grained analysis of explicit and implicit sentiment in financial news articles”, Expert systems with applications, Elsevier, Vol. 42, No. 11, July 2015.

[16]. A Balahur, JM Hermida, A Montoyo,” Detecting implicit expressions of emotion in text: A comparative analysis, Decision support systems, Elsevier,Vol. 53, No. 4, Nov, 2012.

[17]. MS Akhtar, D Gupta, A Ekbal,” Feature selection and ensemble construction: A two step method for aspect based sentiment analysis”,Knowledge based systems, Elsevir, Vol. 125, June 2017.

[18]. Priyanka Patil, Pratibha Yalaagi,” Sentiment Analysis levels and techniques: A Survey”, IJIET, Vol.6, No. 4, Apr,2016.

[19]. DIH Farias, P Rosso,” Irony, sarcasm, and sentiment analysis”, Sentiment analysis in social networks, Elsevier, 2017.

[20]. J Zhu, H Wang, BK Tsou, M Zhu,” Multi-aspect opinion polling from textual reviews”, 18th ACM conference on Information and Knowledge management, Nov, 2019.

[21].Davd Vilares, Carlos Gomez, M A Alonso,” Universal, unsupervised (rule-based), uncovered sentiment analysis, Knowledge Based Systems, Elsevier, Vol.118, Feb,2017.

[22]. R prabowo, M Thelwall,” Sentiment analysis: A combined approach”, Journal of Informatics, Vol.3, No. 2, Apr,2009.

[23] G Grefenstette, P Tapanainen,” what is word, what is sentence, Problems ok Tokenization”, researchgate, 1994.

[24]. AS Zharmagambetov,AA Pak, “ Sentiment analysis of a document using deep learning approach and decision trees”, ICECCO, Sep,2015.

[25]. K Dhana sree, ”Data Analytics: Role of activation function in Neural Net”, IJITEE, Vol.8,No.5,March, 2019.

[26]. C Dos Santos, M Gatti,” Deep convolutional neural networks for sentiment analysis of short texts”, proceedings of COLING, 2014.

[27] B Duncan, Y Zhang, “ Neural networks for sentiment analysis on Twitter”, IEEE ICCICC