Opinion Mining and thought Pattern Classification with Natural Language Processing (NLP) Tools

Sayyada Muntaha Azim Naqvi[1], Muhammad Awais[2], Muhammad Yahya Saeed[3], Muhammad Mohsin Ashraf[4]
Dept. Software Engineering
GCUF
Faisalabad, Pakistan

***Abstract*—Opinion mining from digital media is becoming the easiest way to obtain trivial aspects of the thinking trends. Currently, there exists no hard and fast modeling or classification over this for any society or global community. The marketing companies are currently relying on sentiment analysis for their products. In this paper social sentiment is focused on the form of collective sentiment and individual sentiment; we intend to classify these in the form of Macro and Micro-social sentiment. The sentiment varies among groups, sects etc. and various classes of society are depending on many other characteristics of the society. The social media is available to explore certain ideas, various trends, and their significance. The significance requires further exploration of more patterns and this cycle continues. The exploration cycle focuses on a research outcome. Based on above all the study focuses on the opinion classes towards the general think patterns. The Think Patterns (TP) are developed over time due to social traditions, fashions, family norms etc. The specific community think patterns are very difficult to classify like a female in restricted societies or rural societies of our country. Such trends and patterns are the focus of this study based on various defined parameters. The opinion and sentiment data analysis will be assessed using natural language processing (NLP) tools, Twitter, GATE, Google API's, etc.**

Keywords*—Opinion mining; sentiment analysis; natural language processing; think pattern; GATE

I. INTRODUCTION

The opinion refers to the processes that lead to decisions, such as political, marketing or purchasing decision. Here, the question is which opinion has an influence, whether it is liberal and individual or controlled by power processes. The Internet supports us in different ways to thrive in all units of industries. All Social media like Twitter, Facebook, LinkedIn, YouTube, Myspace, and many others have won a lot of repute that they could not be overlooked[1]. The Internet deals with efficient ways of communicating and distributing opinion.

People express their opinions in the form of natural language. Opinion mining (OM) is one of the natural language tools to track the mood of the audience about a particular product, either it is negative, positive or neutral[2]. Opinion mining is also recognized as the Sentiment analysis. It aims to control the relationship of the author to measure against working, towards some topic or the overall contextual polarity of an article.

Currently, the internet has provided open access to a vast number of texts that are accumulated, like on specialized feedback sites, social networks, and blogs, in the comments sections of news publications. The automatic detection of text sentiment can be used to solve many important applications, the search in a commercial organization regarding consumers to its products, the development of a recommendation system for buyers of certain groups of goods or services, and the introduction in the human-computer interface. The computer system function is responsible for the adaptation of the behavior of the system to the real emotional state of a person etc.

The difficulty of automatic text analysis to determine the emotional relations is expressed in English terminology as Sentiment Analysis (SA). Opinion Mining was among the active scientific research in the early 21st century [9]-[10].

Sentiment analysis is also beneficial for a lot of important applications, like, the research for a commercial organization of the relations with customers for its production, or the development of a recommendatory system for the customers of specified groups of goods or services. The macro sentiment is the study of the sentiment of the national economy as a whole; on the other hand, micro sentiment includes individual, groups or company level. The micro sentiment is done with controlling units oppositely; the macro has unit production. In macro sentiment, stuff is made, and in micro sentiment, stuff is used.

The opinion of other people has huge influences on our behavior, beliefs, and perspective of the world from which we make choices. Therefore, when it is necessary to make a decision, we are often interested in the opinions of others. Opinions are important not only for individuals but also for organizations. Automatic recognition of opinions in texts finds application in a variety of areas: in marketing research, advisory and search systems, in the human-machine interface, in assessing the sentiment of news, etc. [3]-[4]. One of the main errands in the analysis of opinions is the classification of text by sentiment analysis. The tone of the text is the emotional evaluation of some object, determined by the totality of the constituent text of lexical units [5] and the rules of their combination.

The marketing companies are currently relying on sentiment analysis for their products. However, such trends are equally important for industrial and non-industrial users. The opinion and sentiment mining are interrelated fields but unfortunately considered same and focused carefully in our study to clarify it. The research on our community on this

issue is an uprising need of the time to obtain the potential usage of this growing field.

Opinion mining is beneficial for social media monitoring because it permits us to have a general idea of social opinion about some topics. The use of opinion mining is very wide and influential. The capacity to mine ideas from the social data is a way which is extensively used by organizations round the globe. The purpose of the research of opinion is to classify social trends based on moods, opinions, hopes, attitudes, and anticipations of the public or stakeholder groups. In recent years, social networking has transformed into relational communication.

Recent research on language analysis in social networks has focused on its impact on our daily lives, both professional and personal. Natural language processing (NLP) is one of the best favorable approaches to data processing and social networks. NLP is a tool that can assist to run your business progress by providing a visual modality in the minds of the focused audience. However, it is not aimed to change the intuition of a person. There are two different types of a component of NLP: Natural Language understanding and Natural language generation. NLP emphasizes six steps Lexical Analysis, Pragmatic Analysis, Entity extraction, Discourse Integration, Semantic Analysis, and Syntactic Analysis.

Identifying TP is a modest method to understand complex circumstances and develop simply to collect a better understanding of the complex situations and complex covering of a surface of interaction patterns that provoke, drive and direct them [6]. TP helps to understand complex circumstances and develop simple operations to transform them.

The terminologies knowledge discovery, data mining machine learning and pattern recognition in databanks are difficult to detach, using them mainly overlay in their scope. Pattern Recognition is an ability to recognize a set of data regularities, repetitions, similarities or regularities. This feature of the higher cognitive system is being researched for the human perception of cognitive sciences such as perceptual psychology, for machines, however, by computer science. A typical example of the countless application areas is speech recognition, text recognition and face recognition, tasks that are constantly and easily done by human perception. However, the elementary ability of classification is also the cornerstone of conceptualization, abstraction and inductive thinking, and ultimately of intelligence so that pattern recognition has also gained central importance for more general areas such as Artificial intelligence or data mining.

Internet users do not always write constructive and structured feedback on goods or services, considering in detail the pros and cons, exposing the estimates. Much more often the user leaves a spontaneous emotional response in social networks or micro blogging. Further, the spread of smartphones contributes to the increase in momentary feedback and emotional notes on social networks. Like a person after watching a movie, if he did not truly like this, by using a smartphone at the spot without leaving the cinema can warn his friends that it is not worth spending time on this movie. Short notes are written more often and potentially have a greater impact on the friends of the user than the unfolded reviews of strangers. Therefore, for goods, services, media persons and significant events, it is important to collect.

All available collections in Pakistan are collections of reviews belonging to one particular subject area, but not general collections of short texts (micro blogs) or messages from social networks. Therefore, for the task of classifying texts from social networks by tone, a corpus of short texts was built by the micro blogging platform Twitter.

Twitter is a Social network and a micro-blogging facility that permits consumers to inscribe posts in real time. Frequently the message is written from the mobile device directly from the scene, which adds a message of emotionality. Due to platform limitations, the length of the Twitter message does not surpass 140 letters of alphabet. About this feature of the facility, brief messages are issued in real time; individuals practice abbreviates words, spelling miscalculations, smileys etc. Since Twitter has characteristics of a social network, its consumers can actively formulate their view on a diversity of topics from the characteristic of multivariate to the political and economic events in the world.

Classification at the level of expressions and short phrases, instead of whole documents or paragraphs, was conducted by Hoffmann and Wilson, Wiebe [7]. In their work, the writers revealed that it is significant to ascertain the color negative or positive of a particular sentence, not the entire manuscript. In a lengthy article, the writer's view concerning the object be able to replace from negative to positive and positive to negative; the writer can negatively express regarding slight deficiencies, although in general, it remains positive about the object. Furthermore, it is not always possible to classify a long document or a review as positive or negatively colored.

II. Related Work

Vinodhini & Chandrasekaran (2012) explain that Natural Language Processing (NLP) is the domain of sentimental analysis used for tracking the opinion of people on public issues, products, news, and information. The sentimental analysis is also termed opinion mining including the system to gather the views of the public on different blogs of social media (Facebook, Twitter, Instagram, etc.)[1]. In a distinguished manner, the sentimental analysis can be utilized by the users. For example, in marketing, it benefits in mediating the success of the company product.

Zhong et al. (2012) told us that numerous data mining methods had been suggested for beneficial mining arrangements in documentations containing text. In spite of this, how to adequately utilize them and refresh patterns remains a subject of research, particularly in the field of text mining. Meanwhile, the greater part of the techniques for the scholarly investigation of the content implemented term-based methodologies,[2] they experience the difficulties of the considerable number of issues of synonymy and polysemy. Throughout the years, individuals regularly happen in the supposition that the pattern- based methodologies are relied

upon to improve the situation than the term-based. However, numerous tests don't affirm this suspicion.

Unnisa, M., et al. (2016) explained that Social media is one of the most important media for expressing opinions. Analysis of sentiment is a method through which the information is taken as of the Feedback report and the feelings of people about the organizations, actions, and their characteristics. SA also recognized as the OM.

Data discover their tactic to social networking websites such as LinkedIn, Twitter, and Facebook. Twitter provides a wide manifesto to predict consumer brands, movie Critics, democratic elections, the stock market, [3] and the admiration of Celebrities. The key objective of SA is to cluster the negative and positive effects on a bunch of tweets.

III. Methodology

We use GATE in our research. GATE is a platform for deploying and developing software constituents, which process like natural language. GATE is used for performing different tasks. In our research, we make corpora of different documents using the power of Gazetteer. Corpora are the collection of different documents. Gazetteer concerns to recognize the objects name within the content formed on the lists.

For identifying the thinking pattern, as shown in Fig. 1, we take text from different sources and then classify this text according to its category, e.g., classification 1 ton. The text passes through NLP tool then define the Metadata of these categories and implement this Metadata into a different classification. We get Dynamic corpus as the outcome of the above procedure. We repeat this again and again, and as a result, our corpus becomes strong and also identifies trivial patterns.

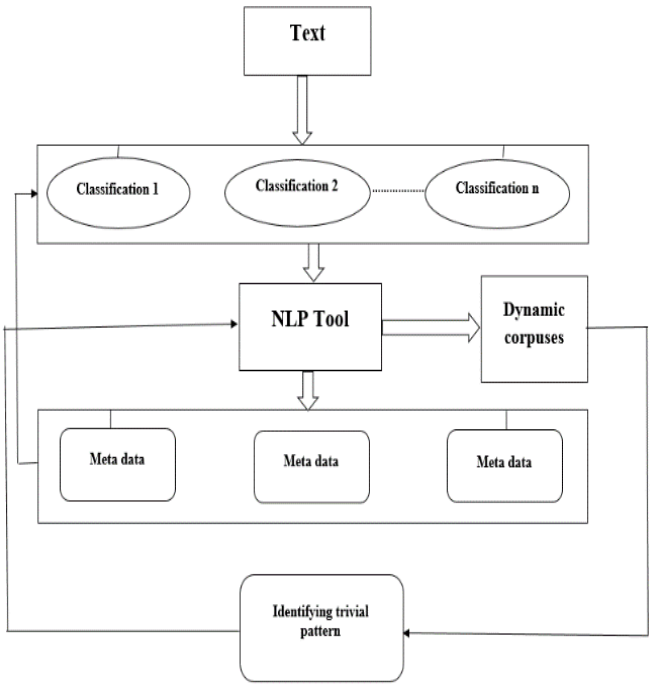

Fig. 1. Exploration Cycle.

Variables:

P: various patterns

S: Sources

T: Various Texts

C: A particular Classification

Crp: corpus

TP: Trivial Patterns

Identifying Pattern

$$p€ Pi \text{ where } 1 <= i <= n \quad (1)$$

Get text from different sources

$$s€ Sj \text{ where } 1 <= j <= n \quad (2)$$

Depends on Patterns and Sources

$$Tq£ PxS;\ T \quad (3)$$

Divide it into various classification

$$Crp: NLP\ () \quad (4)$$

=>If dynamic corpora make

$$Crp => TP \quad (5)$$

Else obtain metadata and classify again

As Fig 2 demonstrate the flowchart of processing. First of all potentially relevant documents or corpora are identified. The information is retrieved from these documents. These corpora are turned into a machine-readable format so that data can be extracted. The meaningful information is extracted and mined to discover new knowledge. When the information is retrieved and normalized the documents, then textual analysis and entity identification is formed. In the end, the required information is extracted, and knowledge is acquired.

The sentiment of the text is determined by the calculation of the weights of the appraisal words included in it. For each text T from the training collection, two weights are counted, the first of which is equal to the sum of the positive evaluation words, and the second is to the sum of the negative evaluation words:

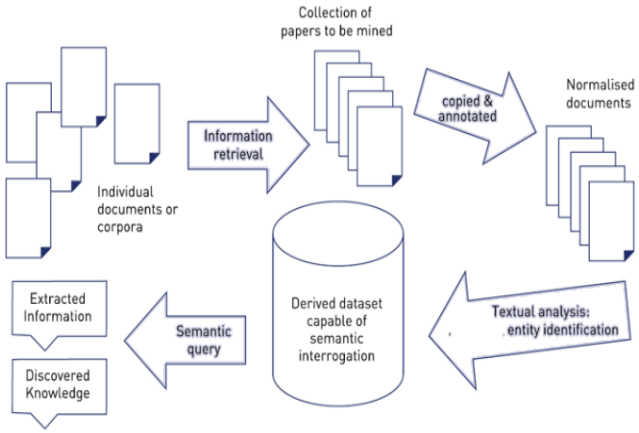

Fig. 2. Processing of Information Extraction [8].

$$W_T^C = \sum_{i=1}^{N_C} |W_i| \quad (6)$$

Where W_T^C is the weight of the text T for the key C; wi is the weight of the estimated word i; NC - the number of evaluation words for the key C in the text T.

All texts of Ti are placed in a two-dimensional evaluation space (positive sentiment - negative key) by their weights W_T^C. To classify texts by sentiment, a linear function

$$f(WT\ pos, WT\ neg) = WT\ pos + k\ neg * WT\ neg \quad (7)$$

Where $WT\ pos$ is the positive weight of the text T; $WT\ neg$ the negative weight of text T; the $k\ neg$ is a coefficient that compensates for the predominance of positive vocabulary in a speech [9]. If the value of the function f is greater than zero, the text is positive, otherwise - negative.

In this research, our goal is to mine opinion of different people as well as targeting general thinking patterns. The general framework of our research is given below as can be seen in Fig. 3.

First of all, we make corpora of different documents. Data analysis is done in the next phase, and processing phase starts after the analysis and specification of the data. NLP tool GATE is used to make the Gazetteer.

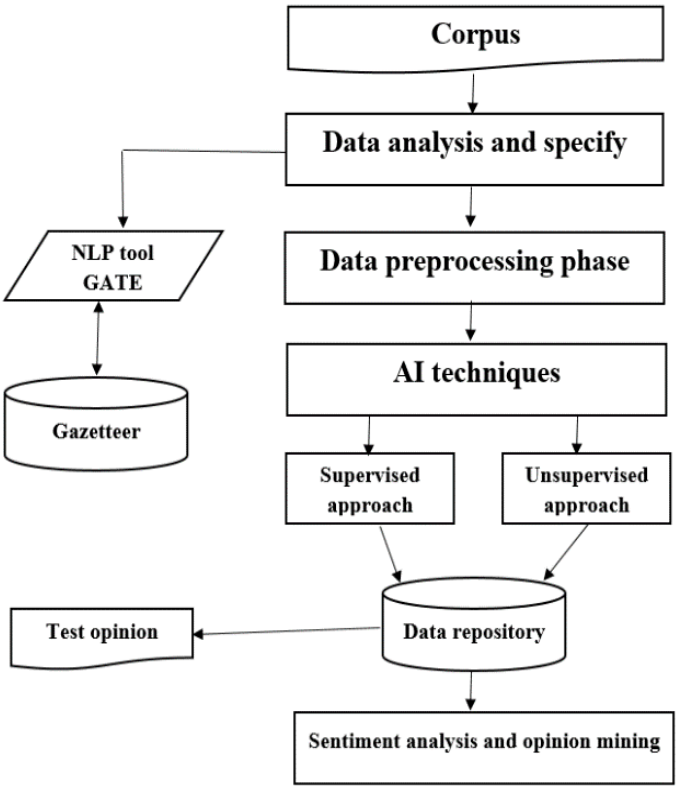

Fig. 3. Framework.

After the preprocessing phase, supervised and unsupervised artificial intelligence techniques are used to retrieve data and enter it into the data repository. By using the data of the repository, we can easily find out the sentiment as well as mining of the opinions. We make corpus that can be related to education, marketing and we can easily mine opinions from these corpora. We find out annotations from these corpora and relate these annotations with education, marketing, and society. Now we have a predefined set of annotations. When these annotations match with that data in corpora, then it is identified that this data supports or relates to that specific area or field. This corpus is used for content filtering and content resemblance. When we want to find the potential category of the document then the content is added into the corpus, it matches to the gazetteer, and we can easily find the potential category of the document, and according to that category, we can mine the opinion. We have corpora which have a large amount of data and go through filtering.

IV. Methodological Issues

A. Making the Relative Corpus

For making different corpus choose Language Resources New GATE corpus, and after that, we have to specify the name of the corpus. In the above pic, it is shown that by using GATE we make corpora of different documents (Fig. 4).

B. Adding Related Data to the Corpus

In corpora, a different type of data is given according to our requirements. We are going to make Education, Business, Teaching, politics, and E-Governess related corpora see Fig. 5. The related data is added to the related corpora [10]. For example, education-related all information is added to the corpus of Education.

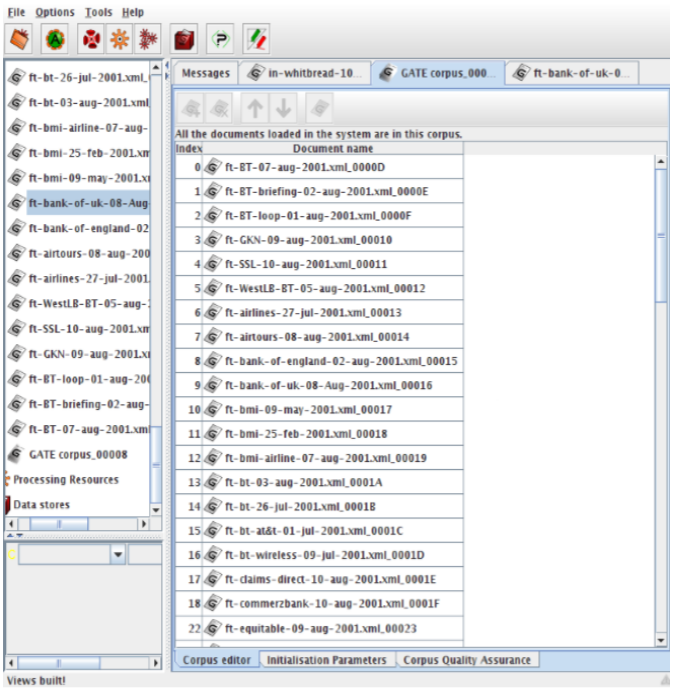

Fig. 4. Different Corpuses.

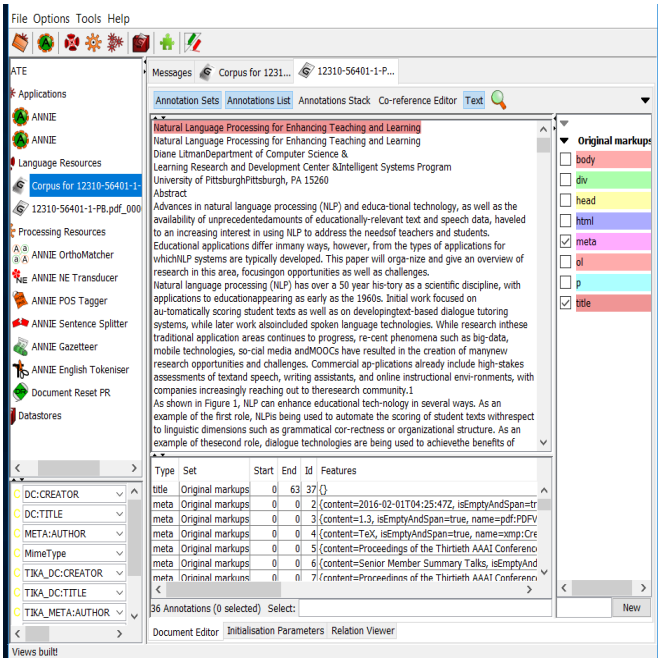

Fig. 5. Adding Data to Corpus.

C. Making Multiple Files for Multiple Types of Gazetteers

A Gazetteer is used for creating annotations. Gazetteer has a large number of sets of lists consisting names of entities like days of the week, cities, organizations, etc. This kind of lists is used to find an occurrence of these names in text. We can add multiple gazetteer processing resources to the controller one gazetteer per list. Def file. The index file particularly references a list file. List files reference major Type, terms, minor Type, list of display name and language. The definition (index) file can comprise list file references. We can insert all lists in one def. File along with each list containing a unique label. A small set of the def list can construct with a related list and also construct a large set of def. Files that change from other lists (see Fig. 6).

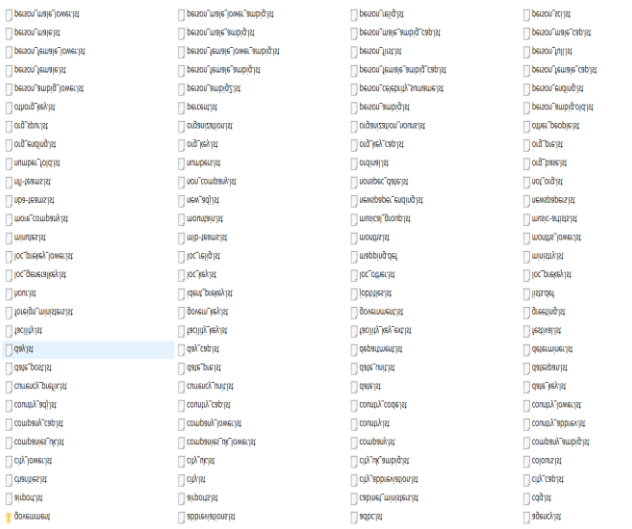

Fig. 6. Multiple Types of Gazetteer Lists.

D. Processing Gazetteers for the Vulnerable Annotation

The double meaning terms have global binding with certain annotations. The intentions of searching text improve with proper use of permutations. The positive and negative permutations as per the search can be classified and improved by training the corpora and jape files.

We have to take different annotations and then sort out the positive and negative annotation. We bring forth a list of positive annotations as well as negative and another list is also created which have contained discarded annotations. For instance, we have the combination of two words X and Y and these two words make the sentiment in the negative direction, and if the major type is X and it is a negative and minor type is Y and Y is positive, then the combination of both will never be highlighted. These types of combination automatically moved towards the rejected list. For illustration, we have a paragraph, and it contains these words runs, scorers and wicket and these words match to our annotation, so it implies to Cricket. We match terms and apply sampling techniques that if these types of words come then, we can identify the topic of the potential text.

We map our words to each other Like, if we have a word and it has negative aspect then we take the second word which is thoroughly positive and combine this word to that word so the results go clear and by intermixing the terms we can change or reduced the content. When we introduced the positive terms, then the search probability and the resolutions will increase and much improved. In most of the cases when one article and one property combines then the results will always be confident. The ambiguous terms which have no proper meaning are combined with the positive terms to improve our search results as shown in Table 1.

TABLE I. ANNOTATION LIST

Meaningless words	Meaningful words
Find Locate Big Small …… …… ……	Children Books Games School ……. …… …….

E. Annotation List and Sets

An annotation set comprises of all individual annotations made by one reviewer or author for course or documents. One annotation set is accessible at a time, but a document can have more than one annotation set. Every annotation inside a set of annotations is related with only one element in a page or a document.

When the set of annotation is created, it automatically creates annotations inside the set see Fig. 7. When we run corpus double click on the loaded application then choose Run and as the resulting corpus that was loaded will be automatically annotated.

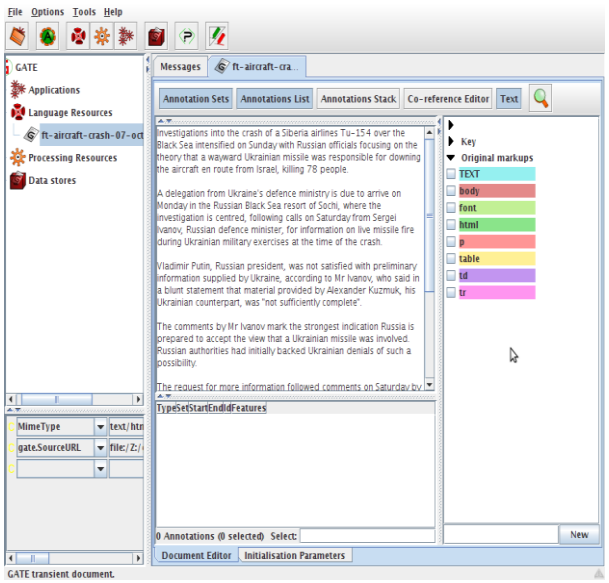

Fig. 7. List and Sets of Annotation.

F. Taking Annotations from the Internet

Whaley believes that the annotation will be a Central part of the future of the Internet, and he is working hard to make this vision a reality. Annotation is positive as well as negative. Every interesting topic or agenda has a lot of annotations.

Global annotations or major annotations are made by avoiding the community. The search engine makes annotations, and it makes annotations on the bases of the search of users. Annotation clears the aspects of objects. Annotations are the small notes that help us to keep track of necessary events like goals, necessary mentions, marketing campaigns, downtime of the website, sales promotions, and time specific events and changes in them. They help us to understand the trends, spears in traffic, and uncommon variations. Google annotations are taken as an example. There are two types of Google annotations shared and private. Every person accesses shared annotations while on the other hand, private annotations are only available for one person. Annotations permit you to note down a specific event that may have an impact on your data Creating SEO based corpus.

If we want to do SEO through NLP, then we have to take those annotations which have higher CPC (cost per click). The benefit of SEO based corpora is that we can dominate our business online. Marketing experts in SEO choose the most reliable keywords for sites to obtain a higher ranking.

Capturing the local domain specific annotation

In the local server, we monitor the URLs. We make a Gazetteer of URLs and compare the URLs. In this, we compare the common URLs with search URLs. In search URL the terms come with the + sign and we can perform aging by extracting these terms and database is used in this process.

G. Uses of AI Techniques

1) Unsupervised approach: Unsupervised learning is employed to mine adjacent blocks of text from a raw stream of characters as the main logical units of an item.

2) Supervised approach: Supervised learning is used to categorize the blocks of different Meta category data, including authors and associations. Afterward, a heuristic is applied to identify the section of references at the termination of the article and break the link to the series [9]. The sorting order used for sorting the tokens of separate links to further information like the reference year and journal. As a final point, we use named entity recognition methods to get links to funding agencies, research grants, and EU projects.

H. Making the Data Repositories and Dynamic Corpus

A repository is a central place where data is stored and managed. A data repository refers to an enterprise data storage entity into which data has been specifically partitioned for an analytical or reporting purpose. The data repository is mostly worked inter-changeably with a data warehouse or mart. It is chosen when a certain type of data storage entity is not identified or is irrelevant to the context [7]. A data repository intends to keep hold of a certain population of data separated so that it can be mined for greater awareness or business intelligence or to be utilized for a specific reporting prerequisite. When data is stored in repositories, dynamic corpora of different type of information is automatically formed.

I. Use of Jape Files

JAPE stands for Java Annotation Patterns Engine. JAPE gives limited state transduction upon annotation established given general articulations.

JAPE enables you to perceive general articulations in explanations on archives. The Jape punctuation comprises of a set of segments, every one of which comprises of an arrangement of activity rules. The segments proceed consecutively and establish a cascade of finite state transducers on annotations. The left half of the rules contain portrays of annotation format (Fig. 8). The right-hand side comprises annotation control statements. The explanations comparing to the LHS of the principal can be determined on the RHS by utilizing the names that are joined to design components.

```
Phase:     DateHeader
Input: DCT
Options: control = appelt

Rule: DCT
(
 {DCT}
):tag
-->
{
gate.AnnotationSet tagSet = (gate.AnnotationSet)bindings.get("tag");
gate.Annotation tagAnn = (gate.Annotation)tagSet.iterator().next();

gate.FeatureMap features = Factory.newFeatureMap();


String s = gate.Utils.stringFor(doc, tagAnn);
//String content =
doc.getContent().getContent(tagAnn.getStartNode().getOffset(),
 //               tagAnn.getEndNode().getOffset()).toString();

 if (s.matches("^\\d{8}$") ) {
String s1 = s.substring(0,4) + "-" + s.substring(4,6) + "-" +
s.substring(6,8);

doc.getFeatures().put("document-date", s1);
}

 }
```

Fig. 8. Jape Files Representation.

J. Treat Web as a Mega Corpus

The web offers new conceivable outcomes for information gathering

- Feasible source of the expendable corpus, constructed ad hoc specially appointed for a particular objective
- Important to work with particular dialects

It is Essential to extract web corpus information while manual mining is time-consuming automatically. Web as the corpus is reasonable for a person who reads with little experience and little learning of semantics, corpus, yet further experienced corpus can discover a few gems in there too [11].

As shown in Fig. 9 Sentences are broken down into tokens and after that language is identified. When a language is identified, translate that language and discover the sentiment of sentences either positive or negative or neutral.

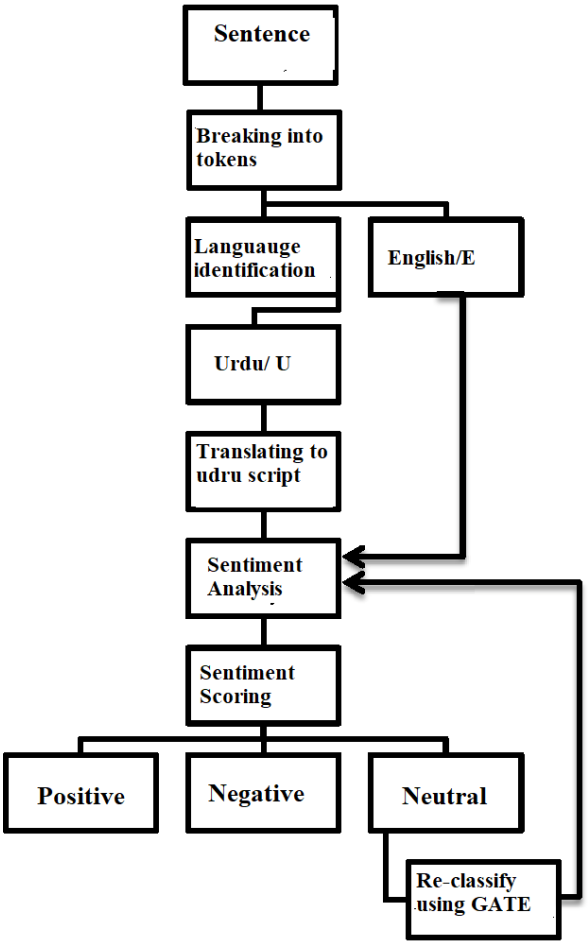

Fig. 9. Flowchart.

V. Application Recommendation

1) Traffic monitoring: Network traffic monitoring is the process of investigating, examining, and processing network traffic or deviations, or a procedure that can influence the performance of network, accessibility, and additional security. We can monitor the traffic of data by using IP addresses, Protocol, and customer parameters.

Analyze the results, and it can be shown in graphical forms or in the form of tables, which is helpful to monitor the real-time usage of internet reporting. In traffic monitoring of internet data, we can check the bandwidth and check the functionality. This traffic monitoring is helpful for commercial as well as personal use. Fig 10 shows the traffic monitoring of GCUF.

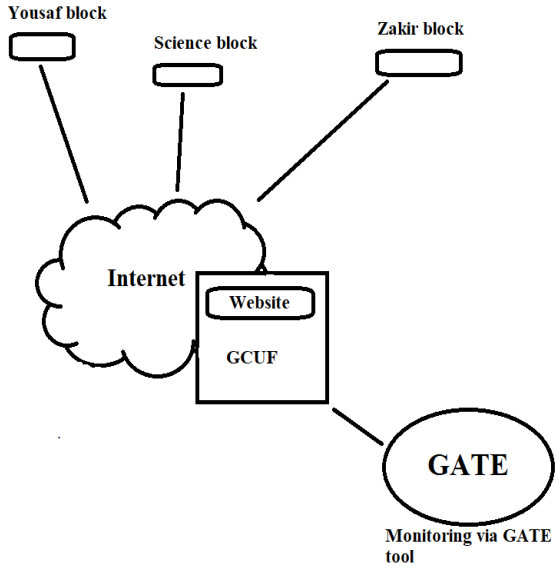

Fig. 10. Traffic Monitoring.

2) URL corpus utilizing for Annotations: People give their URL, and we capture their trends according to the given URLs. People search against URLs. In our server computer 20,000 URLs passes through a day. We make corpora of these twenty thousand URLs, and by making the corpora of these URLs, we get a lot of annotations. By utilizing every annotation, we analyze these annotations.

3) Archived official data processing: It can be maintaining the record the government offices. Like, in a government organization, many workers work in their domains, and they all have their, raw data in a soft-form. We have to take this data to all the employees and make e corpora of all the data, and by making corpora we get an official corpus, and it can also process the records of government offices. Only by entering the required information it can fetch all the useful information which is needed.

4) Local educational annotation identification: Copy the URLs and content of students from the main site, over the computer and make corpora of these URLs. We can know the URL which is used most; frequently we can check this URL either it is related to the educational purpose or not. We can monitor the activities of students by checking their URLs based corpora.

5) General business data analysis: A business firm has a lot of raw documents and emails. We can make the corpus of these emails and process these emails and documents. When we have a lot of corpora, then any time when a business

person needs to analyze their data can use these corpora and analyze their data according to their requirements.

6) Timeline study: Timeline demonstrates the sequence of events from the first to the last alongside the line. This makes clear to understand when the entities have happened compared to another event. Time-lines moreover assist you to study the period closer.

In timeline study, the differences of textual patterns are checked like in a corpus one document is added recently, and other documents stored five to ten years ago, we can easily check the textual pattern difference in this report.

7) Textual patterns identification: Different methods are developed for textual pattern identification. The patterns can be viewed as a semantic and syntactic pattern. Identification of textual patterns means that understanding the meaning of the sentence.

Nowadays a lot of information in government, business, institutions, and industries are stored in the form of text in the database, and this database encompasses a large amount of unstructured data. By using Data mining, we can find out the patterns from a huge database [12]. It can also help us to find out the patterns or correlate between lots of fields in the large database. A pattern is called knowledge if it is interesting and certain enough according to the criteria or measures of the user. A system may face a problem when identified or discovered patterns are not interesting to a user. Similar textual patterns also identified. Some work, have different textual patterns but we can find out similar patterns of these words like, we have three words people, society and culture. They have different textual patterns, and we can identify the similar textual patterns of these words by holding in the annotations of these words (Fig. 11).

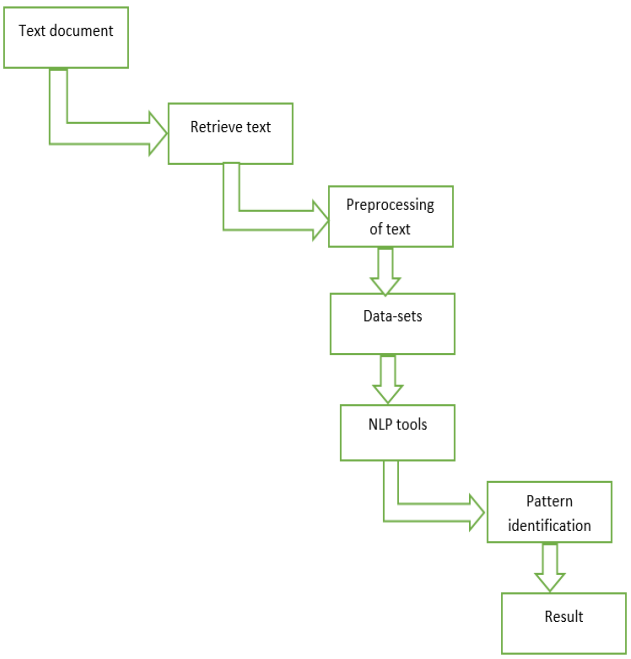

Fig. 11. Identification of Patterns.

8) Improvement in decision making: It can help in, decision making. In our opinion mining research, the sentiment mining, think pattern mining and emotion mining is done all together, and these are helpful in decision mining. We can mine the thinking patterns of the different community. We can mine the opinion and the sentiment of people about a product, and this can be helpful in decision making either we have to change our product or not because we have meaningful information about our product that what people think about the product either they like or dislikes.

VI. Summary

In our opinion mining research, the sentiment mining, think pattern mining and emotion mining is done, and these all together are helpful in decision mining. It can also be helpful to find out the patterns or correlations between a lot of fields in the large database.

If someone desires to make his/her marketing ad, he/she can stimulate the analysis of NLP from us, that what type of keywords he uses, which are the most popular keywords at the same time in marketing. Many marketing companies depend upon SA for their products. GATE is utilized in our research study. Exploration cycle is focused as the outcome. Weights of the words can easily find out using the algorithm. NLP is a very effective approach for multiple domains like teaching, business, E-commerce, politics, and community health. We practice NLP in our research. NLP supports a machine to read the text and also helps to translate it into the natural language in understandable human form. NLP techniques are becoming more common day by day, and these methods are using a lot of data from the internet. NLP is used for information abstraction, classification of sentiment, recognition of E-mail, segmentation of text, language training, and automatic translation, etc. It has two major methods of machine learning and statistical inference [9]. We use GATE (general architecture for text engineering) tool for our research. It is a tool for developing software components which process like natural language. It processed annotations and text directly and parsed this text. We implement Parts of speech; make different lists of the gazetteer. I make a corpus of the university, government school, teaching, industry and also make multiple corpora (corpus within the corpora). I use the power of Gazetteer on the corpora. Gazetteer list is a list of lookups of entities which gathered different files that helps to identify the annotations. I make different corpora, and in these corpora, I find out multiple annotations and add related data to the corpora, and after that, I make multiple files for multiple types of gazetteers and process this gazetteer for vulnerable annotations.

In my research, I take lots of annotations from the internet. I take Google annotations which helps us to understand the trends, monitor the traffic and variations among different trends. I also make SEO based corpora which is very helpful for businesses for getting a higher ranking. I compare the URLs by making gazetteer.

AI techniques supervised, and unsupervised approaches also used to extract the information. Dynamic corpus is generated from the data repositories. We can also make the

corpora of websites. We treat the web as a mega corpus and extract the data which we require. Traffic monitoring of the network is done. We can monitor the traffic by using IP addresses and protocol of the network. The record of government offices is also maintained by making different corpora of their data. We can identify the local education annotation and also analyze the data of the business. The patterns of similar text and identification of patterns are performed by checking the annotations of words and documents. We can mine the opinion and also classify the thought pattern by using a natural language processing tool GATE.

REFERENCES

[1] Bandorski D, Kurniawan N, Baltes P, Hoeltgen R, Hecker M, Stunder D and Keuchel M 2016 Contraindications for video capsule endoscopy World J. Gastroenterol. 22 9898–908

[2] Liu B 2012 Sentiment Analysis and Opinion Mining Morgan & Claypool Publishers Lang. Arts Discip. 167

[3] Junaid S M, Jaffry S W, Yousaf M M, Aslam L and Sarwar S 2017 Sentiment Analysis and Opinion Mining - A Facebook Posts and Comments Analyzer 22 98–104

[4] Kasthuri S, Jayasimman L and Jebaseeli A N 2016 Procedure of Opinion Mining and Sentiment Analysis Techniques: A Survey Mach. Learn. 573–5

[5] Zhang L, Ghosh R, Dekhil M, Hsu M and Liu B for Twitter Sentiment Analysis

[6] Pipanmaekaporn L and Li Y 2012 A pattern discovery model for effective text mining Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 7376 LNAI 540–54

[7] Hertveldt K, Robben J and Volckaert G 2006 Whole genome phage display selects for proline-rich Boi polypeptides against Bem1p Biotechnol. Lett. 28 1233–9

[8] JISC 2012 The {Value} and benefits of text mining Jisc

[9] Chiappe L M 2010 Enantiornithine (Aves) Tarsometatarsi and the Avian Affinities of the Late Cretaceous Avisauridae Author (s): Luis M . Chiappe Published by : The Society of Vertebrate Paleontology Stable URL : http://www.jstor.org/stable/4523457 ENANTIORNITHINE (AVE Society 12 344–50

[10] Anon citation-309438853

[11] Dictionary C S 2012 Methodology for Text Classification using Manually Created

[12] Vinodhini G and Chandrasekaran R 2012 LR... fig 2...Sentiment Analysis and Opinion Mining: A Survey Int. J. Adv. Res. Comput. Sci. Softw. Eng. 2 282–92