

Temporal Sentiment Analysis Using Temporal Synset And Metadata

A PROJECT REPORT

*Submitted in partial fulfillment of the requirement for the award of the
degree of*

MASTER OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Under the Supervision of

Dr. RAJNI MOHANA

(Assistant Professor)

By

Akanksha Puri

(152206)



JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, SOLAN – 173234,

HIMACHAL PRADESH

MAY, 2017



JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

(Established by H.P. State Legislative vide Act No. 14 of 2002)
P.O. Wahnaghat, Teh. Kandaghat, Distt. Solan - 173234 (H.P.) INDIA


Website: www.juit.ac.in
Phone No. (91) 01792-257999
Fax: +91-01792-245362

CERTIFICATE

This is to certify that thesis report entitled “**TEMPORAL SENTIMENT ANALYSIS USING TEMPORAL SYNSET AND METADATA**”, submitted by Akanksha Puri in partial fulfillment for the award of degree of Master of Technology in Computer Science & Engineering to Jaypee University of Information Technology, Wahnaghat, Solan has been made under my supervision.

This synopsis has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Date: May 2017


Signature

Supervisor's Name: Dr. Rajni Mohana

Designation: Assistant Professor

Computer science & Engineering

JUIT, Solan

Acknowledgement

The great desire to acquire higher qualifications and pursue research drove me to promise my parents, brother a useful research work. My promise was to live up to their expectations and never let them down. And I kept my word.

I express my heartfelt gratitude to all those who have contributed directly or indirectly towards obtaining my master's degree and at the same time I cherish the years spent in the department of Computer Science and Engineering Department. I am extremely indebted to my esteemed supervisor **Dr. Rajni Mohana**, who has guided me through thick and thin. The project would not have been possible without her guidance and active support. I am indebted to **Dr. S.P. Ghrera** (H.O.D of Computer Science and Engineering) for providing all kinds of the facilities to carry out the research. Also I want to thank to our coordinator **Dr. Pardeep Kumar** for their inspiration and guidance.

I would like to thank my parents Mr. Rakesh Puri and Mrs. Sunita Puri, and my brother Aman Puri who have supported and encouraged me at every stage of my life. Without them, I would never have had neither the confidence nor the tenacity to do the research. They had been a constant source of inspiration to me.

Date: May-2017


Akanksha Puri

152206

List of Contents

| | | |
|------------------|--|----|
| Chapter 1 | Introduction | 1 |
| 1.1 | Natural language processing | 1 |
| 1.2 | Sentiment analysis | 3 |
| 1.2.1 | Components based on sentiment analysis | 3 |
| 1.2.2 | Sentiment analysis uses | 7 |
| 1.3 | Temporality | 7 |
| 1.3.1 | Temporal sentiment analysis | 8 |
| 1.3.2 | Expressions of documents and their normalization | 10 |
| 1.3.3 | Applications of temporal sentiment analysis | 11 |
| 1.3.4 | Temporal tagging | 12 |
| 1.4 | Frequently used terms | 13 |
| Chapter 2 | Literature Survey | 14 |
| Chapter 3 | Problem Description | 19 |
| | Proposed Solution | 21 |
| Chapter 4 | Methodology | 24 |
| Chapter 5 | Implementation and Results | 26 |
| Chapter 6 | Conclusion | 33 |
| | Future Scope | 34 |
| | List of References | 35 |

List of Tables and Figures

| S.No. | Title | Page No. |
|--------------|---|-----------------|
| 1. | Table 1 Comparison Between various Research Papers | 16 |
| 2. | Figure 1 Components of building sentiment analyzer | 4 |
| 3. | Figure 2 Levels of abstraction | 4 |
| 4. | Figure 3 Featured levels of abstraction | 5 |
| 5. | Figure 4 Knowledge for knowledge extraction | 6 |
| 6. | Figure 5 Types of temporal sentiment analysis | 10 |
| 7. | Figure 6 The basic process of text classification | 24 |
| 8. | Figure 7 Model for implementation of temporal tagging | 26 |
| 9. | Figure 8 Front Page of Model | 27 |
| 10 | Figure 9 Pre-processing of data | 27 |
| 11 | Figure 10 Main design of page | 28 |
| 12 | Figure 11 Uploaded data | 28 |
| 13 | Noun list and Verb list | 29 |

Abstract

Natural language processing is a most common area of research that probe how computer understand and manipulate natural language text or speech to do useful things. Sentiment analysis is one of the active research area in natural language processing. Sentiment analysis is process of determining the emotional tone behind a series of words, distinguishing and classifying opinions expressed in an article, in order to achieve whether the writer's attitude towards a areas in natural language processing. Sentiment analysis based on temporality is gaining much attention in many real time applications. This manuscript highlights the key concepts of various state-of-art temporal sentiment analysis along with the research gaps. It also put focus on the normalization of various temporal expressions. It covers different temporal expressions and the methods for normalization of these expressions. The main focus is to find the temporal tagging of data. To facilitate the future work, a discussion of state-of-art resources and methods for temporal sentiment analysis is also provided.

CHAPTER-1

Introduction

1.1 Natural Language Processing:

Natural language processing is a field of computer science and artificial intelligence appertained with the interactions between computer and natural languages. It is sometimes also known as an AI-complete problem. Natural language processing is a most common area of research that probe how computer understand and manipulate natural language text or speech to do useful things ^[1].

NLP is a procedure in which user analyze the computer, understand it and extract meaning from human language in a useful way. Using the natural language processing, many tasks can be solved such as sentiment analysis, automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.

Applications of natural language processing

Natural language processing, is a branch in which human languages become easy to understand by computer. In coming years, Natural Language Processing will become a main equipment to traverse the space between human communication and digital data. There are five ways that natural language processing will be used in the years to come.

1) Machine Translation

As the time changes rapidly, there is very fast progress in technologies that are used. Today's information about everything is available online. The job of making that information available becomes ever more important. The challenge of making the information available, across language limit, has simply outgoing the ability for human translation. Many large companies are looking to recruit large amounts of people to contribute, translation efforts with learning a new language. In machine translation we have more

appropriate alternate to matching the information available online. There is a challenge with machine translation technologies that it cannot translate the words, but it secure the meaning of sentence.

2) Fighting Spam

Spam filters have become imperative as the first line of security in defense against the increasing problem of unwanted email. But now everyone uses email service and has experienced anguish over unwanted emails that are received by them, or important emails that have been inadvertently caught in the filter. The false positive and false negative issues of spam filters are at the spirit of NLP technology. There is a filtering spam technique that is widely used is Bayesian spam filtering, this is a arithmetical technique in which the frequency of words in an email is calculated against its normal occurrence in a quantity of spam and non-spam emails.

3) Information Extraction

In financial markets many important decisions are increasingly affecting away from human supervision and control. Algorithmic trading is a form of financial investing that is entirely composed by technology becoming more popular in the world. But many of these financial decisions are crash by news, journalism which is still presented mainly in English. NLP performs a major task in this, taking plain text and extracting the relevant information in a format that can be deal into algorithmic trading decisions.

4) Summarization

Overloading information is a real fact in our digital time, and our right to use information and capacity to understand it. This is a way which shows that there is no signal of slow down, and has the capacity to summarize the importance of documents and information is becoming ever more important. This is very useful for us as it allow us to recognize and understand the relevant information from vast amount of data available on net. Another preferred result is to understand wide emotional meanings of information, like analysis of sentiment on any product. This application of NLP will become more useful as precious marketing benefits.

Natural Language Processing is used for the text analysis, allowing computer to understand human languages. The way how human-computer interacts facilitate real-world values. NLP is generally used for text mining, machine translation, and automated question answering.

NLP is considered as a hard problem in computer science. Human language is rarely precise or plainly spoken. Human language can be understood by knowing the concept and information about it how they are linked and its meaning.

1.2 Sentiment Analysis:

There are many tasks that come under natural language processing like sentiment analysis, mane-entity-recognition, translation, summarization and many more. Sentiment analysis is solitary of the dynamic research area in natural language processing and content mining. Opinion mining is another name given to Sentiment analysis. Sentiment analysis is process of determining the emotional tone behind a series of words, distinguishing and categorizing opinions expressed in a entity, in sort to derive whether the writer's thoughts towards a areas in natural language processing.

The effort in Sentiment analysis is a group of text engineering or Deep learning. Deep learning is a system which translates the machine what they recognize. It can be defined as different sentiments are processed to support automatic decision formation. Prior dispersion of sentiments into binary form was the major task of SA, i.e., to classify them into positive and negative sentiments. Modification in the trim of granularity was multiply along with the technical exploration in the machine learning.

1.2.1 Components of building sentiment analyzer:

The various steps to build a sentiment analyses is shown in Figure 1. It express the ways how knowledge is acquired from different sources, then classify the level of abstraction and then by using any suitable method extraction process is done and then update the knowledge base for further.

Figure 1: Components of building Sentiment Analyzer

- 1) *Knowledge acquisition:* It can be defined as to take out real world sentiments through existing algorithms, it is essential to train the analyzer for learning about real world entities and sentiments. The archive of such a massive knowledge is called as knowledge base. It is the basic element of sentiment analyser. This knowledge base can be constitute using information from different sources like linguistic artists, lexicons, books, newspapers, magazines, articles etc.
- 2) *Selection of level of abstraction:* In sentiment analysis there are three different levels of abstraction i.e., document level, sentence level and feature level as shown in figure 2.

Figure 2: Levels of abstraction

- a) *Document level:* Document level is the top most stage of abstraction in sentiment analysis. At this level the binary classification of text document is done. It works on the coarse grain level for the extraction of granularity. Document level sentiment analysis gather the largely positive and negative opinion of high requirement.

- b) *Sentence level*: Document level and sentence level processes of sentiment are almost the same origination processes. At sentence level, its main task is to detect the documents subjectivity. As Sentence level sentiment analysis means subjectivity detection, so there is increase in the granularity which is more than document level. In this the text document holds sentences whose nature can be categorized as subjective and objective. An opinion analysis depends on the knowledge of individual's about any entity or article. The main task at this level is to consider only subjective sentences and does not include the objective sentences.
- c) *Feature level*: This level is the premier level of granularity. In this document can be classified as implicit or explicit expressions. First they are classified into implicit or explicit then further divided it into domain independent and domain dependent.

Figure3: Featured level of abstraction

3) *Knowledge Extraction*: It is the part of analyser, in which by using various existing methods and learning techniques the resultant of all the problems are collected on the basis of pre-extent knowledge base. Any model can be used for extraction process there is no hard rule for using them.

- Models to approach the huge collection of data which at times is very useful and sometimes meaningless, there is a need of machine-learning to originate it. There are choices of machine-learning models used by sentiment analyzer to mine meaningful

data from large information. It is very necessary to choose appropriate model. There is no hard rule on the selection of models. When a wrong model gives a correct prediction then that model is treated as a useful model.

Figure 4: Model for knowledge extraction

Models are categorized in the following types:

- i) Predictive models understand the future value of the query in the topic. These models may call as today for future (TF) models. The efficiency of the models used for the calculation depends on its result, i.e., the correctness of predictive values. Naive Bayes, (Pang et al., 2002) ARM, (Pang et al., 2002) etc. methods are used to predict the future values for the output.
- ii) Descriptive models (Liu, 2012) are in reality used to summate the analysis. In sentiment analysis there are various techniques that are used for the descriptive models such as Classification, clustering and ARM etc.
- iii) User-sensitive models (Liu, 2012) of impression are used to follow the fact that for the one user group same opinion could be positive and on other hand negative for another. E.g. For a photographer a camera can be rated as a good, but too difficult for the casual user.

iv) Author authority models description for the fact that personal authority often significance whether the reader will be influenced (Liu, 2012) significantly by a given review or not.

1.2.2 Sentiment analysis uses:

Sentiment analysis is quite useful in social media survey as it grant us to get an overview of the wider public opinion behind definite topics. Social media monitoring tools like Brand watch Analytics make that process faster and easier than ever before, thanks to real-time monitoring capabilities.

- The applications of sentiment analysis are extensive and prevailing. The ability to select observation from social data is a practice that is being widely supported by various groups across the world.
- Any change in the opinion or attitude on social media can affect the changes in stock market.
- The capability to rapidly understand the customer attitudes and act in response accordingly to them is used for the sentiment analysis.
- Sentiment analysis is used by the various brands, companies, organizations for their use. Many brand and companies uses sentiment for the feedback of their product whether it is good or bad. They fulfill the needs of user or not.

1.3 Temporality:

In the recent years, temporal tagging has acknowledged rising study in the area of natural language. Temporal tagging is an area of Sentiment analysis along with natural language. Temporality is the state of actual within or having some relationship with time. It is traditionally the linear sequence of past, present and future. It plays a vital role in sentiment analysis to calculate sentiment strength of any entity. Temporal sentiment analysis ^[2] is Aggregate sentiment communicate at different points in time and perform trend analysis by looking at how sentiment changes over time.

1.3.1 Temporal Sentiment Analysis:

Sentiment analysis is the part of Natural Language Processing, in which the opinion of user is detected and classification of attitudes in texts are involved [22]. For the classification of sentiment, different paths or methods can be used such as supervised learning, semi-supervised learning and unsupervised learning.

- Supervised learning is the part of data mining used to train the system. It is used for the labeled training data to train the system. The training dataset consists of a set of various training examples. There are various machine learning algorithms that are used for supervised learning of data like SVM, HMM, etc. In 2002, Pang et al. describes Naive Bayes, maximum entropy and SVM different approaches for analysis of sentiment to categorize reviews into positive and negative opinions. In supervised learning each example or training set consists of pair values of an object like input value and its output value. Attribute extraction in this is done on the basis of grouping of all active titled individual.
- Semi-supervised learning designs are used for the inputs which are the combination of labeled and unlabelled data. In 2004, Pang and Lee describes Graph-based semi-supervised learning methods which are based on minimum cuts. Etzioni et al. Proposed bootstrapping method is also proposed by for the minimization of manual labeling of input data.
- Unsupervised learning techniques are used to deal with unlabelled or defined data analysis. In unsupervised learning the main task is to find hidden composition in unlabelled data. In this rules are frame to train the system with help of unlabelled data.

Nowadays, time is one of the key features that conclude document reliability besides importance, correctness, impartiality and exposure. Temporal expressions are usually anchored by events & attributes of the temporal expressions are useful in distinguish between temporal relation of the events

plus ordering of the events. NLP and Information retrieval indicates different title of analysis for temporality. NLP aims to understand time at a fine grained level. From the IR viewpoint, temporality has been studied at a coarse-grained level.

TempoWordNet is an essential source for time related applications both in NLP and IR. Temporality means having some relationship with time. It is traditionally the linear sequence of past, present and future. In last few years, temporality has accepted increased attention in natural language processing and information retrieval. In temporal information retrieval, text document contains the systematic and algorithmic protocol which involves steps for the creation, extraction and normalization. In text documents two standards are used for annotating temporal information: TIDES TIMEX2 ^[19] and TimeML ^[20]. Both standards immediate guideline for how to verify the extents and how the value of temporal expressions is normalized.

The following attributes are used in annotation for normalization:

VAL: For temporal information it is a normalized form of the expressions in ISO standards

MOD: Modifier of temporal expressions.

ANCHOR VAL: It is a normalized form of an anchoring date or time.

ANCHOR DIR: the relation direction between VAL and ANCHOR VAL.

SET: It is used to identify expressions denoting sets of times.

From these attributes the four types of temporal expressions dates, times, durations and sets can be normalized. The value attribute of date and time expressions directly pertain to a point in time.

For Example: “2016-03-15” for the expression “March 15, 2016”.

For durations and set expressions, it covers the length of the time interval, e.g., “P5D” for “Five days” and “every five days” ^[8]

Thus, temporal sentiment analysis analyzes temporal direction of sentiments and topics from a text document that has timestamps. It is the relation of time with

sentiment. It is useful in detecting the mood of user in different time frames. Temporal sentiment analysis is used in many research areas.

1.3.2 TEMPORAL EXPRESSIONS IN DOCUMENTS AND THEIR NORMALIZATION

TimeML, the standard mark-up language for temporal annotation, we have four types of temporal expressions: dates (Dec 3, 2016), times (7 p.m.), durations (one weeks), and sets (daily).

Types of Temporal Expression

There are four kinds of temporal expressions: Explicit, Implicit, Relative Expression and Underspecified Expression as shown in figure. In temporal tagging each document is associated with a domain that is to be processed and pertain domain specific strategies for extraction and normalization of temporal expressions.

Figure 5: Types of temporal expression

Explicit: In explicit expression, the temporality is captured through the metadata.

For example: *Explicit expression is referred as “25/November/2016” or “25-11-2016”.*

Implicit: In implicit expressions, the temporality of the document is hidden in the document itself. It can be captured by applying some language rules over the text.

For example: *Implicit expression sometimes written as “Children’s Day 2016”.*

Implicit expressions are normalized by using knowledge about their meaning.

Relative: These expressions are those which form a specific relation with a temporal focus.

For example: *Relative expressions may written as “after three days” or “Yesterday”.*

For the normalization of relative expression identification of reference time is required.

Underspecified: It is an expression where users neglect the elements of information, which then have to be improved from the context by using some language rules.

For example: *This expression sometimes written as “in June”.*

To normalize the above expression, it is required to identify the temporal relation to the reference time. In the News and Narrative style document corpora, it is simple to identify the reference time. In these documents the reference time is mostly equals to DCT (Document creation time).^[8]

1.3.3 APPLICATIONS OF TEMPORAL SENTIMENT ANALYSIS

Recently many researchers find that online texts analysis can be helpful for trend or event prediction.

- *Forecasting Political Results:* Political orientations^[16] are presage from the reviews and comments of users. Users post their ideas or opinions about each politician or party over the web sphere.
- *Mood Detection:* In sentiment analysis, time plays an important role in mood detection. Temporal sentiment analysis describes the mood swings of users. This is useful in predicting the behavior of user.
- *Popularity/ Infamy:* In^[17] On the basis of sentiment analysis the achievement of box office is predicted. From the reviews or comments on movie we can rate the movie or drama.

- *Stock Market:* In ^[18] Stock markets movement is read from the news article sentiment or mood of twitter users.

1.3.4 Temponyms Tagging:

For many NLP and IR functions, it is most important to extract temporal information from text documents. Thus, temporal tagging means the extraction and normalization of temporal expressions. In recent years it has attained a lot of interest of people and proposed a number of tools such as HeidelTime and SUTime. Temporal tagging means temporal scopes for textual phrases. In this the detection of temporal expressions and make their semantics accessible by normalizing them into standard format is done. The four types of temporal expressions are used i.e. Explicit, Implicit, Relative and Underspecified. For the tagging, the combination of three data sources is used to create a large collection of explicit temponyms with their temporal values i.e. A) Yago B) Aida C) Temponyms pattern directory.

The Temponyms are created in following steps:

- 1) Prephased all the temporal facts from the Yago using alias names from aida directory. Yago contains all the data like starting on date, finishing date, time when happened, start and end of event.
- 2) Aida has knowledge about alias names for entities.
- 3) Then, create temponyms by matching predicates to noun phrases in pattern directory.

1.4 Frequently Used Terms:

Temporal sentiment analysis is related to time with sentiment. There are various components which are used for analysis of sentiment. Some of the components are given as:

WordNet: WordNet sometimes also known as ontology. WordNet is any of the machine understandable databases of information about words/text for the English language.

It assort words into sets of similar words or words having same meaning called synsets and narrative a number of relations among synonym sets.^[24]

SentiWordNet: It is sentiment lexicon companion sentiment information to each WordNet synset. It is the combination of WordNet and sentiment information.

SentiWordNet categorized each word in synset of WordNet into three categories: Positive Score, Negative Score and Objectivity Score.

TempoWordNet: It is a set of time-perspective synsets. TempoWordNet is a free verbal knowledge base for temporal analysis where each synset has its own intrinsic temporal value.

TempoWordNet classify each synset of WordNet into four categories: Atemporal, past, present and future.^[25]

Metadata: Metadata is data that depicts other data. Metadata is structured information that delineates, explain, detect, easier to extract and retrieve, use, manage an information. We use metadata for discovering useful and relevant information about each resource.

CHAPTER-2

Literature Survey

In a natural language processing, many researchers work on Temporal Sentiment analysis. Tomohiro et al. ^[3] worked on temporal sentiment analysis and proposed two graph methods: Topic graph and Sentiment graph. They have analyzed the temporal trend of sentiment and topic from text with timestamps. Yoonjung et al. ^[4] described a framework for sentiment analysis for generating context driven features. They proposed methods for in co-operate two tasks: sentiment generation and Sentiment classification. The classification system consists of two parts: contextual feature generation and domain-specific sentiment classifier construction. In this paper, they classified a candidate words into positive, negative and neutral category by using bootstrapping method. Some researchers put efforts to linking text sentiment to public opinion time series. Brendan et al. ^[5] proposed various methods for analysis of text. They measured the public opinion derived from polls with sentiment measured with text from many popular micro blogging sites.

Now, many researchers or scholars start determining the temporal sentiment analysis on social media, they effectively use twitter information for their work. Mike et al. ^[6] determined the sentiment in twitter events. They described three approaches for sentiment analysis in twitter: Full-text machine learning, Lexicon-based methods and Linguistic analysis. They had given some methods for the time series of online topics. They had checked the strength of sentiment. Jannik et al. ^[7] analyzed the sentiment or text on different domains. The authors had given the two types of standards: TIMEX2 and TimeML. They used TIMEX2 attributes for normalized the four type of temporal expressions: DATES, TIME, DURATION, SETS. They described some strategies to handle with challenges of temporal tagging on different domain and their integration with temporal tagger Heidel Time. They had created colloquial and scientific corpus and compare them with news and narrative. Tadahiko et al. ^[8] proposed a system for visualizing twitter users based on temporal changes in impressions. They proposed a web application system for visualizing twitter. In system when user login his account, the system collects the tweets of user that posted during the time period. They identified the impression of users from each tweet and make line and pie chart corresponding to impressions and their temporal change.

Andre et al. ^[9] described methods for spatial and temporal sentiment analysis and sentiment polarity. They implemented two approaches for sentiment classification of tweets and compare their results: 1) SVM 2) Naïve Bayes. They perform

classification process in two steps. In first step only tweets with subjective texts are classified. In second step each tweet is classified with a single sentiment either positive or negative. For classification of sentiment polarity they used Emotions and manual labeling approaches. Gael et al. ^[10] proposed method to build a temporal ontology which contributes to time related applications. They had used Synsets and categorized them into present, past and future. In this they had used the two step classification and one step classification method for categorization. Yuanyuan et al. ^[11] worked on mapping of dense Geo-tweets and web pages. They had designed a tweet mapping system which support web and twitter user communication. In this system matching their location names and categories each tweet based on their category name of floors from web page with their different time frames. They had described many mapping functions such as acquisition of tweets, filtering, acquisition of web pages and clustering of tweets.

Luciano et al. ^[12] discussed about the happiness level of people in a city. They measured the sentiment expressions in tweet that are posted from popular areas in city. They had given many methods for human mobility and sentiment analysis. Syed et al. ^[13] done work on Localized twitter opinion mining using sentiment analysis and had discussed a methodology which allows utilization and interpretation of twitter data to determine public opinions. They had analyzed the gender of user. Now, some researchers start working on Temponyms tagging. Edral et al. ^[14] describes the integration of wide range temponyms. They used temporal tagger to cover subset of temponyms. The authors had detected temporal expressions and make their sentiments accessible by normalizing them into standard formats. They had used four temporal expressions: Explicit, Implicit, Relative and Underspecified. For tagging they used three data sources to create large collection of explicit temponyms with their temporal values: 1) Yago 2) Aida 3) Temponyms pattern directory.

TABLE I. COMPARISON BETWEEN RELATED WORKS

| Ref No. | Paper Name | Work Done | Limitation |
|----------------|------------------------------|---|----------------------------------|
| 1. | Sentiment in Twitter Events. | <ul style="list-style-type: none"> Authors used three approaches for sentiment | 1 month include only two special |

| | | | |
|----|--|---|---|
| | | <p>analysis:Full-text machine learning, Lexicon-based and linguistic analysis.</p> <ul style="list-style-type: none"> • They classify the text as positive or negative. | <p>events the Oscars and the Olympics.</p> <p>The sentiment strength algorithm used is also an issue.</p> |
| 2. | TempoWordNet for sentence time tagging. | <ul style="list-style-type: none"> • The authors had used temporal synset having 21 sets and categorized the sentences into past, present future tense. • They used two classification method and one classification method for analyse the word as temporal and atemporal. | <p>one-step process shows incapacity to solve the temporality issue.</p> |
| 3. | Dynamic Mapping of Dense Geo-Tweets and Web Pages based on Spatio-Temporal Analysis. | <ul style="list-style-type: none"> • In this tweets are related to its category name. They have used various mapping function such as acquisition of tweets, tweet filtering and acquisition of web pages. • Various methods for categorization of tweets: • K-NN method,Naive bayes method,Support vector machine. <p>RMSE and average score value of tweets are calculated</p> | <p>Users cannot obtain the most recent information, while they browse web pages since these are not updated in real time.</p> |
| 4. | Temporal tagging on different domains: Challenges,strategies and Gold standards. | <ul style="list-style-type: none"> • They had created colloquial and scientific corpus and compare them with news and narrative. • They analysed the temporal tagging on different domains and ddressing the diffrent challenges. | |

| | | | |
|----|---|--|--|
| | | | |
| 5. | Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events. | <ul style="list-style-type: none"> Temporal sentiment analysis: The method takes texts with timestamps such as Weblog and news articles as input, and creates two types of graphs: <ul style="list-style-type: none"> a)topic graph b) sentiment graph | They did not proposed method for Weblog articles. |
| 6. | Temponym Tagging: Temporal Scopes for Textual Phrases. | <ul style="list-style-type: none"> They had detected temporal expressions and make their semantics accessible by normalizing them into standard format. For tymponym tagging, they combine three data sources to create a large collection of explicit temponyms with their temporal scopes: YAgo, Aida, Temponym pattern directory. | They only addressed explicit temponyms. |
| 7. | A spatial and temporal sentiment analysis approach applied to twitter microtexts. | <ul style="list-style-type: none"> The author implemented the Classification of Sentiment Polarity by using two approaches and then comparing their results: <ol style="list-style-type: none"> 1.SVM 2.Naive Bayes They detected the geographical location of tweets . | They concerned with the identification of the entity referred to by the opinion detected in tweets |
| 8. | Domain specific sentiment analysis using contextual feature generation | <ul style="list-style-type: none"> Classify a candidate word into positive,negative and neutral category using bootstrapping method. They proposed a domain-specific sentiment analysis | System was not able to capture phrase-level clues. i.e the units of the clues need to be |

| | | | |
|-----|---|--|---|
| | | system utilizing context features in news texts. | expend |
| 9. | Visualizing temporal changes in impresiion from tweets. | They proposed a web application system for visualizing twitter users based on temporal changes in the impressions from the tweets posted by users. | This system is not suitable for the impressions for tweets that used the keywords. |
| 10. | Localized twitter opinion mining using sentiment analysis | <ul style="list-style-type: none"> • The authors had discussed a methodology which allows utilization and interpretation of twitter data to determine public opinions. • They had analyzed the gender of user. | <p>Quality of tweets was also very low.</p> <p>In gender classification there is still some errors.</p> |

CHAPTER-3

Problem Description

Sentiment analysis:

There are many tasks that come under natural language processing like sentiment analysis, name-entity-recognition, translation, summarization and many more. Sentiment analysis is one of the active investigation area in natural language processing and text mining. Opinion mining is another name given to Sentiment analysis. Sentiment analysis is process of determining the emotional tone behind a series of words, distinguishing and assort opinions expressed in a article, in order to derive whether the writer's feelings towards a areas in natural language processing.

Research issues in sentiment analysis:

The prospect of sentiment analysis is to make a system that is essential to have a common sense similarity as human beings. Now this time the system lacking in this state. Today, due to the improvement in the processing task, techniques used in machine learning are also improved. By using various machine learning techniques unstructured data is converted into the structured data. The research gap is to develop a common-sense system for SA of any natural text.

Temporality:

It is the state of existing within or having some relationship with time. It is the linear progression of past, present, and future. Temporal Sentiment Analysis is defined as the aggregate sentiments expressed at varying points in time and perform trend analysis by looking at how sentiment changes over time.

Temporal sentiment analysis analyzes temporal direction of sentiments and topics from a text document that has timestamps. It is the relation of time with sentiment. It is useful in detecting the mood of user in different time frames. Temporal sentiment analysis is used in many research areas.

The research issues of temporal sentiment analysis are following:

Temporal Aspect is another important dimension of Semantic Analysis. Detecting attitude shifts is made possible identifying opinions and identifying when they were stated. Monitoring the impact of marketing campaigns or containing damage to brands and companies through quick response are few of the important applications of temporal dimension. Therefore, need of automatic assignment of high weightage to the recent reviews and low to the previously posted blog, review, etc is required Following are the fields where semantic analysis are used:^[21]

1 Explicit/implicit text document: Most text documents are given explicitly in the query or deduced implicitly. Extraction of topic implicitly is a best but complex option.

2 Geographical distribution of time: The task of sentiment analysis based on the temporal aspect becomes more complex because of deviation in time worldwide. Hence, it is very monotonous to get a collaborative real-time sentiment of the whole world.

3 Forecast analysis: Researchers are working to improve forecast analysis by including the time in the process of sentiment analysis.

4 Hinged weightage to reviews with respect to time: As the time passes, the value of previous reviews is degrading. The present day review is more important. In Sentiment analysis to have time-oriented review the process should include joint weightage with respect to time.

- Our aim is to find temporality of the text using temporal synset and Metadata. To find out the temporal tagging through TempowordNet.

Proposed Solution

To do the temporal tagging of the text using temporal synset and metadata:

In this the tokenization of text document is done. The tokenized text document comes under the process of PoS tagging where each word is classified as noun, verb, adjective etc. Then classify the textual data according to Word Synset, where each word in temporal synset is classified as temporal and atemporal value.

Then categorize each text into their category i.e. past, present and future.

At last, we have rules to find temporal tagged data. Each temporal value has to be compare with the metadata and output is obtained.

For the solution of our task we have used MATLAB software

Techniques used in classification of Text:

Text Mining: Text mining is the process of analysis of information contained in natural language text. Text mining process can use by an organization to collect important and valuable knowledge about any particular topic, business, text documents, E-mails, posts on social media. In text mining removal of unstructured data with natural language processing, statistical modeling and using various machine learning techniques is a challenge, because text in natural language is often unpredictable. In NLP, it contains inconsistent syntax and semantics which cause ambiguity.

Tokenization: Tokenization is the procedure of splitting a stream of text into phrases, terms, signs, vocabulary or other significant elements called tokens. The main aim of the tokenization is the searching for the words in a sentence. Textual data is only a textual explanation or chunk of characters at the starting time. In information retrieval words of the data set are required. So we require a passer which processes the tokenization of the text documents. This may be insignificant as the text is previously stored in machine understandable format. But the removal of punctuations, brackets, hyphens, other symbols etc is still an problem that is not solved. The main use of tokenization is to identify the meaningful keywords. a further problem in this is abbreviations and acronyms which should be changed into a regular form.

Tokenize: This operator divides the text document into a series of tokens. There are numerous options to define the breaking points. The options are as follows:

Mode: In this the selection of mode is done. It chooses the tokenization mode. Build upon the mode, breaking points are chosen in a different way. The choice is non letters, specify characters, regular expression and the default value is non letters.

Characters: The incoming text document will be tear into tokens on each of this characters. For example enter a '!' for split into sentences. The series is string and the default value is '!:'

Expression: This regular expression defines the splitting point.

Stopword Elimination: The most general words that improbable to help in text mining such as prepositions, articles, and pro-nouns can be treated as stop words. Every text document has to deal with these kinds of words which are not necessary for purpose of text mining. All these words are needed to be eliminated. For e.g., “a”, ”is”, “you”, “an”, “and”, “for” etc. we are able to decide any group of word for this purpose. It also cut down the text data and helps in improvement in the performance of system.

Stemming: It is also referred as lemmatization. This is a technique in which words are reduced into their stem, base or roots. There are any words in the English language that can be reduced to their base form or stem e.g. study, studying, studios unlike belong to study. Additionally, names can be changed into root by removing the “s”, for e.g., During the stemming procedure the variation is the word with “s” in a sentence is reduced to normal word and this elimination may show the way to an mistaken stem or root. Although, these words are not used for human communication then, there is no problem in the stemming process. All pronunciation of the root are altered into the same root, so the stem is still helpful.

Filtering: Filtering is the process helps in provide the flexibility when user want to design data sources and mining constitution it provides the flexibility so that on the basis of comprehensive data source view a particular removal construction can be formed. For training and testing part of the different models, only that part of data filters can be formed to use by them, For each subset of data their is no need to construct a different structure. We can use filter by length, Content, English, dictionary and Region etc.

These are parameters used in this operator are:

Min chars: The minimal number of characters that a token must have to be measured.

Max chars: The maximal number of characters that a token must have to be measured.

For the classification process and categorizing them according to their category we have used the MATLAB software. Firstly to find the temporal and atemporal data, we have to classify the data into its category. If the data is classified into past, present and future then it is temporal in nature and if cannot classify to any category then it is atemporal in nature. For the classification process we have used neural network classification method. Neural network is a classification process which is used to train the system. This system is based on training the data, according to which it classify the data into past, present and future. This neural network shows the result on the basis of accuracy.

CHAPTER-4

Methodology

In this first classify the textual data according to Word Synset, where each word in dataset is classified as temporal and atemporal value. Then categorize each temporal synset into their category i.e. past, present and future.

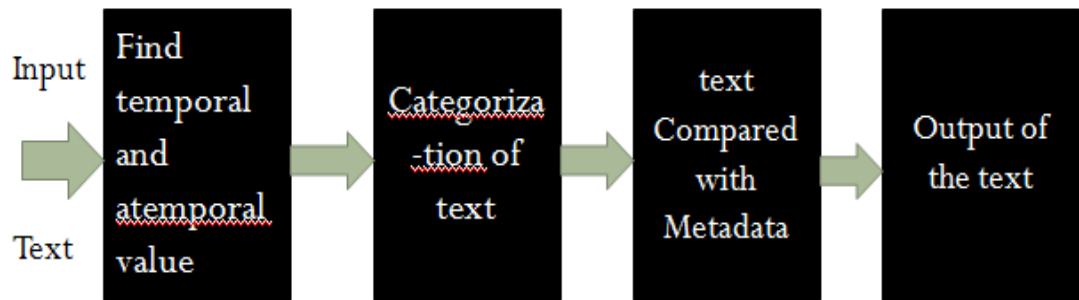


Figure 6: The basic process of text classification

Temporal and Atemporal Classification: When the words in dataset are classified into various classifiers i.e. past, present and future then it is Temporal in nature. If it cannot be classified into any category then it comes under Atemporal classification. The temporal and atemporal values are found from the TempowordNet. The results of the given dataset are compared with TempoWordNet to find out the similar text and find their category to which it belongs. Once the past, present and future classifier has been cultured, there are more temporal categories than past, present and future as there are more than positive and negative classes in the expression of sentiments.

After the classification of text into present, past and future category, each temporal value in the synset has to be compared with the metadata. Each text document has its time of post and date of post through which the data is compared with metadata and obtain their temporal tagged data. We have many rules to compare the data with metadata.

Metadata is the data about data which contains the Date of Post, Time of Post, Year of Post, Place where it is posted etc.

When the data is compared to metadata each document is compared and apply the various rules on it for the classification.

By using different techniques we will classify the text into its category and map each sentence value with metadata to find their improvement and accuracy.

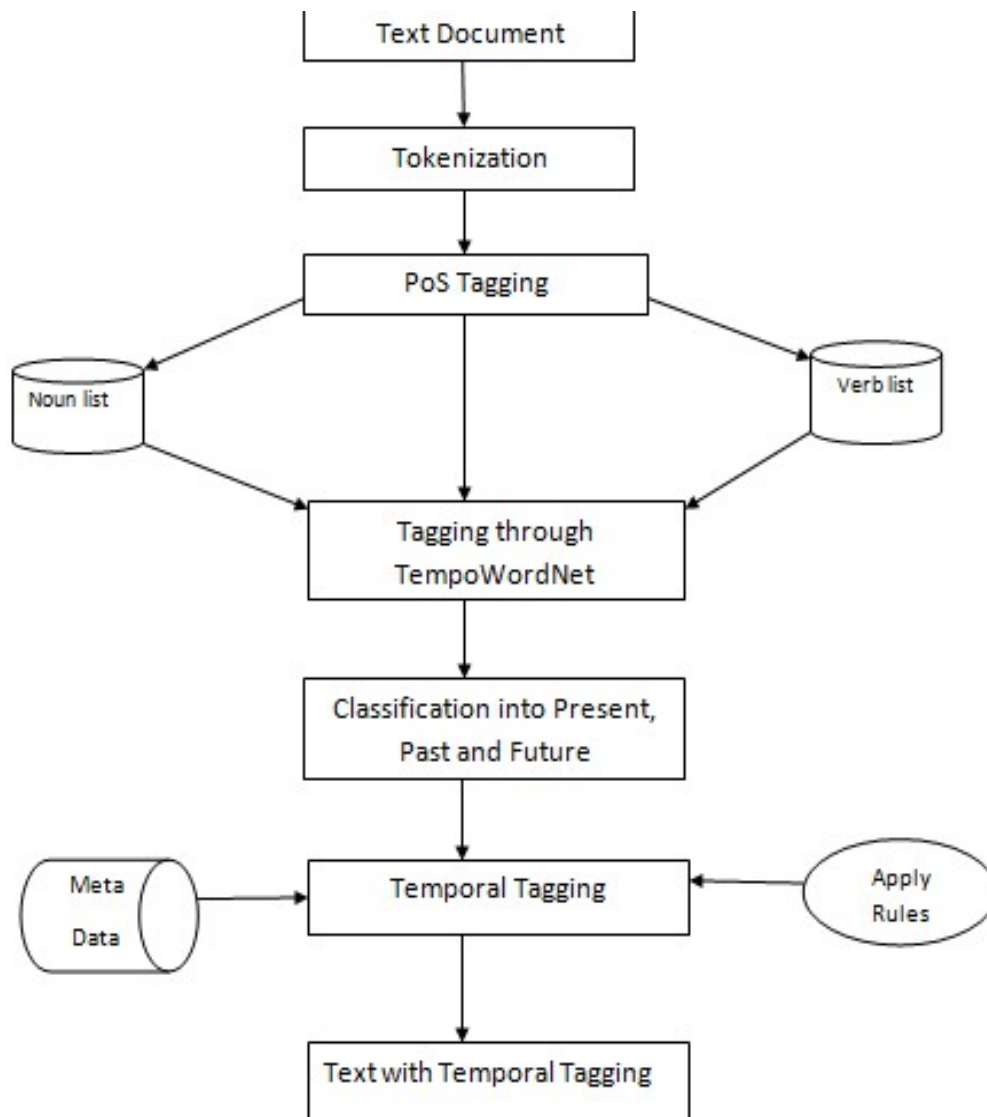
For the classification process and categorizing them according to their category we have used the MATLAB software. First we take the dataset and done pre-processing of data, the obtained result is tokenization of dataset. The tokenized dataset is gone

through its PoS tagging means part of speech tagging in which each word is categorized into verb, noun, adjective and adverbs. In our project we have taken only noun and verb list of the words. The noun list and verb list words are gone under tagging through TempoWordNet. In this step each word is compared with the words in TempoWordNet. In TempoWordNet each word has its probabilistic value according to its category. We have done the classification process according to the nearby belonging category to which the probability value of each word found. For the classification process we have used neural network classifier.

After the classification process is done, we have done the temporal tagging process. In this we compare the results of classification process with its metadata. In our dataset each text has its Time of Post, Date of Post, and Year of post when it is being posted on social media. We have made rules to classify each word with metadata. With the comparison of classified text with metadata we have found its classification with time.

CHAPTER-5

Implementation and Results



Model for the temporal tagging of Text Document

Figure 7: Model for Temporal Tagging

- We have created a user graphics interface for sentence time tagging.

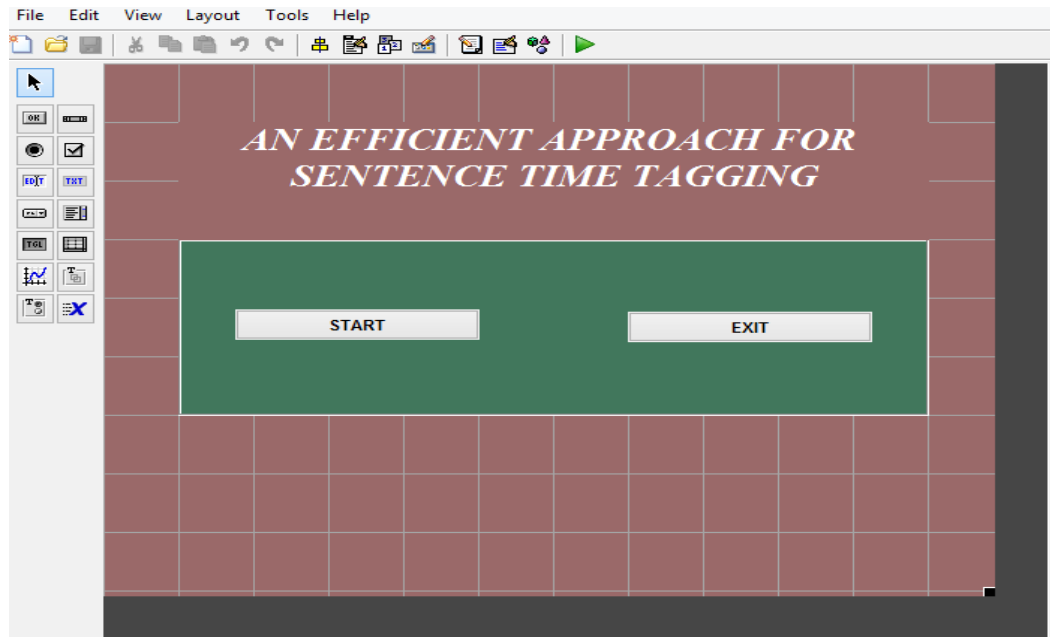


Figure 8: The front page

- *Step-1:*

Tokenization: Tokenization is the procedure of splitting a stream of text into phrases, terms, signs, vocabulary or other significant elements called tokens. The main aim of the tokenization is the searching for the words in a sentence. In our project we have done the tokenization process by doing the pre-processing of dataset. In pre-processing step text document is gone through its pre-processing in which stemming, filtering, all these operations has been done and find the pre processed data in form of different cells.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|--------------|---|-----------|--------------|-----------|------|-----------|---------|
| 1 | @switchfo... | 0 | @Kenichan | @nationwi... | @Kwesidei | Need | @LOLTrish | @Tatiar |
| 2 | | | | | | | | |
| 3 | | | | | | | | |
| 4 | | | | | | | | |
| 5 | | | | | | | | |
| 6 | | | | | | | | |
| 7 | | | | | | | | |

Figure 9 : Pre-Processing of data

After the pre-processing step, stop word has to be removed from the tokenized dataset. We have the list of tokens which does not contain any stop words. The dataset is in the form of cells, each cell contain token of different sentences.

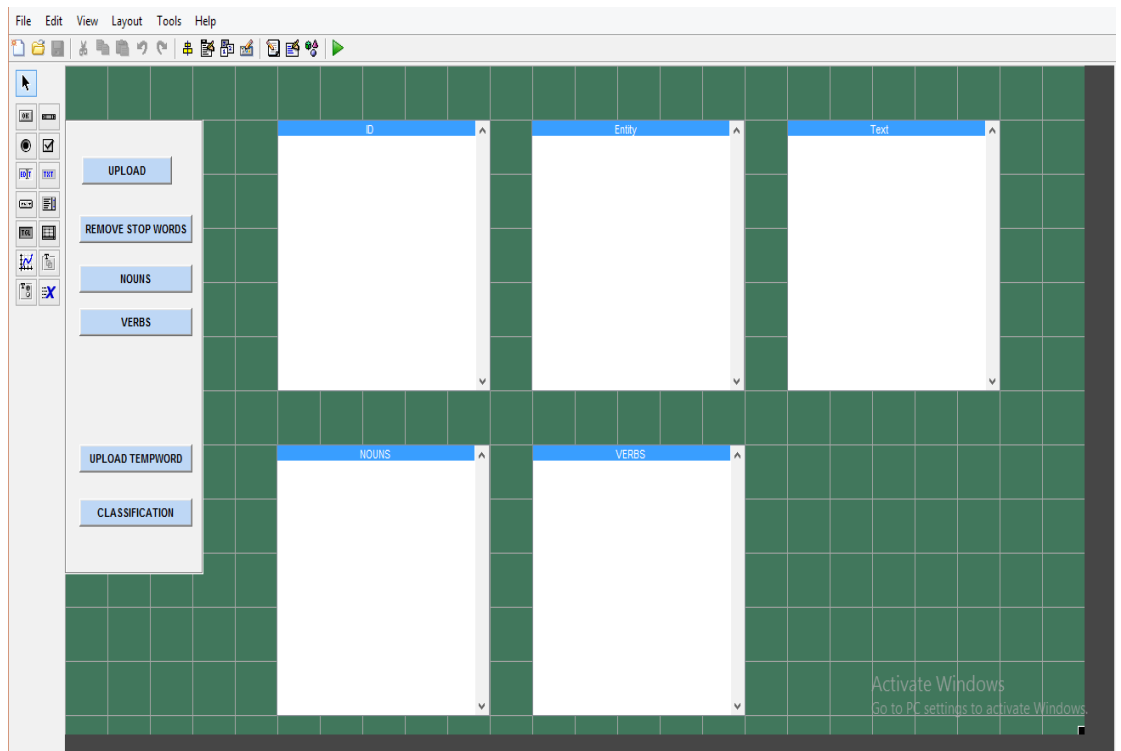


Figure 10: Main design of page

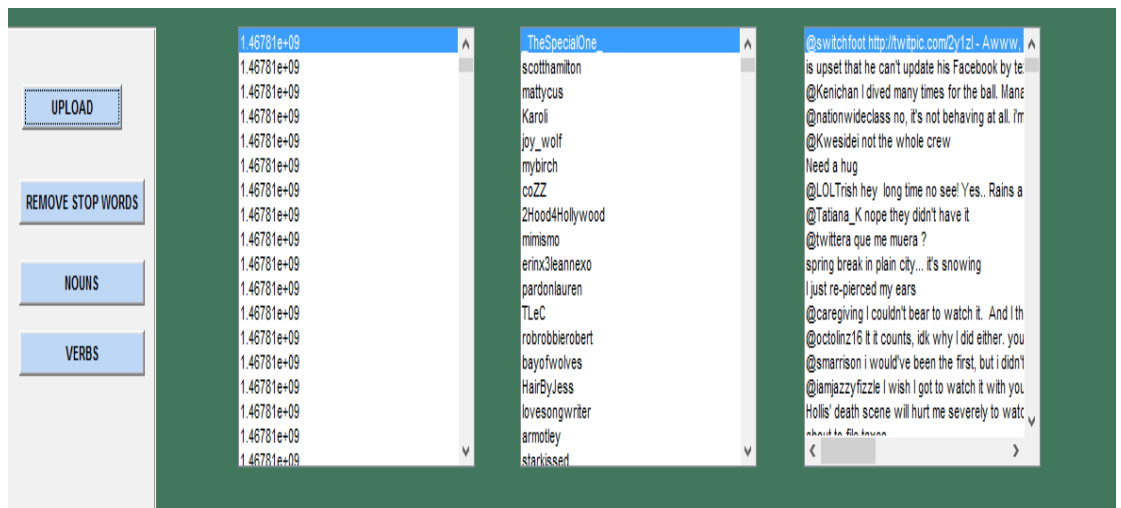


Figure 11: Uploaded Data

- *Step-2:*

PoS Tagging: A Part of Speech is a portion of software that takes the text document in any language and reads it and attaches PoS to each word such as noun, verb, adjective, etc. Each word is assigned to its belonging category.

In this, by doing PoS tagging we categories the text into noun list and verb list. In this model, PoS tagging is done for each token and categories each token into verb, noun, and adverbs.

The obtained dataset is go through its tagging through PoS in which we have classified noun list and verb list of words. Then upload these noun list and verb list in the main frame.

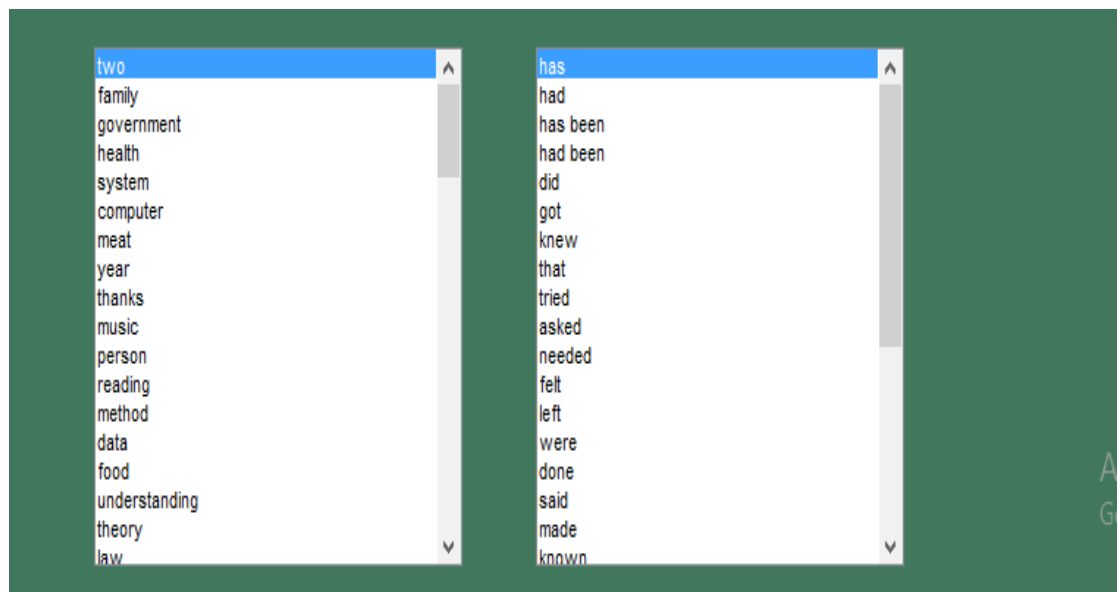


Figure 12: Noun list and Verb list

- *Step-3:*

Tagging through TempoWordNet: It is a set of time-perspective synsets. TempoWordNet is a free verbal knowledge base for temporal analysis where each synset has its own intrinsic temporal value. TempoWordNet classify each synset of WordNet into four categories: Atemporal, past, present and future.

In this step we have done the pattern matching process. In this the dataset contained noun and verb list is compared with the TempoWordNet to find similarity between both of them.

In TempowordNet each word has its belonging category and is classified according to its probabilistic value. Upload the tempowordnet and compare it with existing noun list and verb list.

- *Step-4:*

Classification into Past, Present and Future: By using the TempoWordNet we classify each text into its belonging category such as past, present and future.

- In TempoWordNet each word has its probabilistic values according to its category, to find whether it belongs to past, present, future or atemporal category.

For the classification process we have used Neural network classification method. Neural network is a classification process which is used to train the system. This system is based on training the data, according to which it classify the data into past, present and future. This neural network shows the result on the basis of accuracy.

- *Step-5:*

Temporal Tagging: For the temporal tagging of text we need to apply some rules and metadata. By applying all rules with metadata we have concluded the text with temporal tagging.

Metadata contains the information about data, the date of post, time of post, year of post when the data is uploaded, by using all these values we have compared the results from classification process to metadata and obtain the data with its temporal value.

- In last we have the text with temporal tagging.

Algorithm:

- i. First upload the dataset.
- ii. Pre-processing of data is done and obtains the data in tokenized form.
- iii. Do the PoS tagging of data and find out the noun list and verb list.

- iv. For the tagging through tempowordnet upload the Tempowordnet. Compare it with existing dataset, and find out the similar word between them.
- v. For the classification process by using neural network classify the data into different categories i.e. past, present and future by matching their probabilistic values with tempowordnet.
- vi. The obtained results from classification process are compared through its metadata values and obtain the final results to which category the data belongs.

Performance Matrices:

There are the following metrics that are used to take measure of the system performance:^[15]

To calculate Precision, Recall, F-Measure and Accuracy of the system following performance metrics need to be defined:^[21]

True positives (TP): Number of positive examples classify as positive.

False positives (FP): Number of negative examples classify as positive.

True negatives (TN): Number of negative examples classify as negative.

False negatives (FN): Number of positive examples classify as negative.

1. Precision: The precision is the ratio of the number of relevant documents returned to the total numbers of documents for a given user query.

Precision =

2. Recall: The recall of a text can be defined as the ratio of the number of relevant documents returned to the total number of relevant documents for the user query in the set.

Recall =

3. F-Measure: F-Measure is defined as the measure that combines precision and recall is the melodic mean of precision and recall.

F-Measure =

4. Accuracy: It is defined as the ratio of the number of correctly classified objects to the total number of objects.

Accuracy =

Results:

The percentages of classification of the data into its different categories are:

The probability of being present = 25.48%

The probability of being past = 85.99%

The probability of being future = 44.56%

By comparing the results of classified data to the metadata we obtain the different results which are given:

The probability of being present = 21.93%

The probability of being past = 14.79%

The probability of being future = 40.74%

CHAPTER-6

Conclusion

Temporal sentiment analysis is an evolving area with a variety of real time applications. Although sentiment analysis is a challenging task, much attention is paid to it over the last decade. The rising need of real time applications gives temporality based sentiment analysis more valuable. Normalization makes the processing of web data more efficiently for sentiment analysis. Apart from the state-of-art temporal sentiment analysis, the research gaps described in this manuscript needs to be covered for increasing the performance of the analyzer.

In this we augmented WordNet with temporal information by following a many steps process. Text document is classify as atemporal or temporal and then, all temporal synsets are associated to its past, present and future category. The classification process is done by the neural network classifier which is used to train the system. After the categorization, the data values obtained are mapped to metadata and find their accuracy and compare their results.

Future Scope

In this project we have done the classification process with its temporal values and metadata. In future we will work on the temporal sentiment with high granularity. Temporal sentiment investigates the pattern of sentiment within a given time of interval. We will work on the mood detection of human with change in time.

List of references

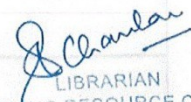
- [1] Chowdhury GG. “Natural language processing.” *Annual review of information science and technology*, 37(1), 51-89, Jan 1 2003
- [2] Preethi PG, Uma V. “Temporal sentiment analysis and causal rules extraction from tweets for event prediction.” *Procedia Computer Science*, 48, 84-89, Dec 31 2015
- [3] Dias GH, Hasanuzzaman M, Ferrari S, Mathet Y. “Tempowordnet for sentence time tagging.” In *Proceedings of the 23rd International Conference on World Wide Web*, Apr 7 2014, pp. 833-838.
- [4] Fukuhara T, Nakagawa H, Nishida T. “Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events.” In *ICWSM 2007*.
- [5] Choi Y, Kim Y, Myaeng SH. “Domain-specific sentiment analysis using contextual feature generation.” In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, 2009 Nov 6, pp. 37-44.
- [6] O'Connor B, Balasubramanyan R, Routledge BR, Smith NA. “From tweets to polls: Linking text sentiment to public opinion time series.” *ICWSM*, 11, 2010 May 23, 122-129.
- [7] Thelwall M, Buckley K, Paltoglou G. “Sentiment in Twitter events.” *Journal of the American Society for Information Science and Technology*, 62(2), 406-418, Feb 1 2011.
- [8] Strötgen J, Gertz M. “Temporal Tagging on Different Domains: Challenges, Strategies, and Gold Standards.” In *LREC*, Vol. 12, pp. 3746-3753, 2012.
- [9] Kumamoto T, Wada H, Suzuki T. “Visualizing Temporal Changes in Impressions from Tweets.” In *Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services*, Dec 4 2014, pp. 116-125.

- [10] Alves AL, de Souza Baptista C, Firmino AA, de Oliveira MG, de Paiva AC. "A Spatial and Temporal Sentiment Analysis Approach Applied to Twitter Microtexts." *Journal of Information and Data Management*, 6(2), 118, Jan 20 2016.
- [11] Wang Y, Yasui G, Kawai Y, Akiyama T, Sumiya K, Ishikawa Y. "Dynamic mapping of dense geo-tweets and web pages based on spatio-temporal analysis." In *Proceedings of the 31st Annual ACM Symposium on Applied Computing 2016 Apr 4* pp. 1170-1173.
- [12] Gallegos L, Lerman K, Huang A, Garcia D. "Geography of Emotion: Where in a City are People Happier?." In *Proceedings of the 25th International Conference Companion on World Wide Web 2016 Apr 11*, pp. 569-574.
- [13] Hridoy SA, Ekram MT, Islam MS, Ahmed F, Rahman RM. "Localized twitter opinion mining using sentiment analysis." *Decision Analytics*, 2(1), 1, Oct 22 2015.
- [14] Kuzey E, Strötgen J, Setty V, Weikum G. "Temponym Tagging: Temporal Scopes for Textual Phrases." In *Proceedings of the 25th International Conference Companion on World Wide Web 2016 Apr 11*, pp. 841-842.
- [15] Junker M, Hoch R, Dengel A. "On the evaluation of document analysis components by recall, precision, and accuracy." In *Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on 1999 Sep 20*, pp. 713-716.
- [16] Park, S., Ko, M., Kim, J., Liu, Y., Song, J. "The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns." in: *Proceedings of the ACM Conference on Computer Supported Cooperative Work. ACM, 2011*, pp. 113–122.
- [17] Schuller, B., Knaup, T. "Learning and knowledge-based sentiment analysis in movie review key excerpts." In: *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues. Springer, 2011*, pp. 448–472.
- [18] Li, X., Xie, H., Chen, L., Wang, J., Deng, X., 2014. "News impact on stock price return via sentiment analysis." *Knowl.-Based Syst*, 69, 14– 23, 2014.

- [19] Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson. "Standard for the Annotation of Temporal Expressions." TIDES, The MITRE Corporation, 2005.
- [20] James Pustejovsky, Robert Knippen, Jessica Littman, and Roser Sauri. "Temporal and Event Information in Natural Language Text." *Language Resources and Evaluation*, 39(2-3), 123–164, 2005.
- [21] Kaur, Sukhnandan, and Rajni Mohana. "A roadmap of sentiment analysis and its research directions." *International Journal of Knowledge and Learning* 10.3, 296-323, 2015.
- [22] Balahur, Alexandra, and Marco Turchi. "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis." *Computer Speech & Language* 28.1 (2014): 56-75.
- [23] Fellbaum, Christine. "WordNet and wordnets." In: Brown, Keith et al. (eds), *Encyclopedia of Languages and Linguistics*, Second Edition, Oxford: Elsevier, 665-670, 2005.
- [24] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." *LREC*. Vol. 10. 2010.
- [25] HULTECH, GREYC- CNRS 6072 Laboratory, Normandie University, Caen, France, January, 28, 2015.

Submission Info

| | |
|------------------|--------------------------|
| SUBMISSION ID | 806858369 |
| SUBMISSION DATE | 29-Apr-2017 12:49 |
| SUBMISSION COUNT | 1 |
| FILE NAME | M.Tech_Akanksha_Puri.... |
| FILE SIZE | 415.41K |
| CHARACTER COUNT | 49487 |
| WORD COUNT | 8943 |
| PAGE COUNT | 43 |
| ORIGINALITY | |
| OVERALL | 24% |
| INTERNET | 15% |
| PUBLICATIONS | 15% |
| STUDENT PAPERS | 80% |
| GRADEMARK | |
| LAST GRADED | N/A |
| COMMENTS | 0 |
| QUICKMARKS | |


LIBRARIAN
LEARNING RESOURCE CENTRE
Jaypee University of Information Technology
Waknaghat, Distt, Solan (Himachal Pradesh)
Pin Code: 173234

