International Journal of Advanced Technology & Engineering Research (IJATER)
www.ijater.com
UGC approved / ISO Certified

AN APPROACH FOR BLLINGUAL SENTIMENT ANALYSIS USING NATURAL LANGUAGE PROCESSING (NLP)

Indrajeet Patel, Research Scholar, Rewa Institute of technology, India
Parikshit Tiwari, Head of Department, Computer Science & Engineering, Rewa Institute of technology, India

Abstract

Deep Learning(DL) is new and important research area of Machine Learning(ML), which works on unsupervised learning methods. As Today, most NLP& ML methods require labeled training data (i.e., supervised learning) but almost all data is unlabeled so processing needs to done by unsupervised learning methods. This issue brought us to use unsupervised Deep Learning (DL) approach. Now a day's much research is going on various Natural Language Processing (NLP) task such as Text extraction, Text summarization, Word Embedding, Sentiment Analysis (SA) and Optical character recognition.

As amount of data on internet is growing by means of internet or by humans and there is a need to extract content from document and retrieve important data from extracted content. Today, most of the implemented systems are using supervised approach. To improve the result accuracy, we are using Deep Learning (DL). The application includes analyzing sentiment of text, representing review, opinion, and polarity of text.

Keywords· Big Data · Classification · Microblogging ·Sentiment Analysis · Social media analyticsMachine-learning approach, Lexicon-based approach, Sentiment classification.

Introduction

NLP is related to human–computer interaction. NLP is a branch of computer science focused on a system through which computer can communicate with humans in natural languages. It is rapidly growing technology at present. However, developing NLP application is bit challenging as it deals with human language, which is not precise and is ambiguous.Common NLP tasks include Part-of-speech tagging, Named Entity Recognition, Text summarization, Co referenceResolution, Sentiment Analysis (SA), Word sense disambiguation. Most of the work in NLP is done using Supervised approach but most of the data is in unstructured form due to this problem we have chosen Deep Learning (DL) approach which follows unsupervised method to perform some of the NLP task.

Our project performs the sentiment anlysis on Hindi and English both language.

1.1 Sentiment Analysis (SA):

Sentiment Analysis (SA) is the task of Natural Language Processing (NLP) and textual analysis to determine the subjective information of text. It is the task of retrieving opinion about products and classifying the review as positive, negative, or neutral. It helps in determining the speaker's attitude with respect to the overall polarity of sentence or document. Sentiment Analysis (SA) is also known as opinion mining.

The main function of Sentiment Analysis (SA) is to classify the polarity of a given text at the document, sentence, or feature/aspect level. It is also use to find the emotions in the sentencesuch as "angry," "sad," and "happy".

For example, "This movie is awesome".

In this sentence, Sentiment Analysis(SA) determines that sentence is about movie and it shows positive review about movie.

In this project, we are extracting polarity and opinion of the sentence or document based on their various feature score. Basis of known data which is training data and that training data consist of input data and response values.

It is a process of categorizing and collecting opinion about the product.

The purpose of Bilingual Sentiment Anlysis using NLP is to fulfill the requirement of extracting the text from unstructured data and performing other Natural Language Processing (NLP) task Sentiment Analysis (SA). Deep Learning (DL) works well on unstructured data and improves accuracy of the results. The amount of unstructured data that humanity produces overall and on the Internet grows, and that data needs logically process it and extract different types of knowledge from it. The reason to follow DL (Deep Learning (DL)) approach is that, nowadays most NLP and ML methods require labeled training data (i.e. supervised learning) but almost 80% of the avalible data is unlabled so processing on it has to be done by unsupervised learning methods. This issue brought us to Deep Learning (DL)approach. Our need is to develop a system, which can perform multiple tasks on text and character recognition to perform Document Image

Analysis that transforms documents in paper format to electronic format.

1.2 Objective:

1. To provide an easy user interface to input text.

2. User should be able to upload the text file.

3. System should display opinion, polarity and emotion of the text.

1.3 Problem Statement:

Extracting text (Hindi and English) from any type of file or web page, review and opinion of that extracted text using Deep Learning(DL)(DL) method to improve the accuracy of the results.

Scope of this project is to extract text(Hindi and English) from different files (PDF, Text, Ms Office, Html) using Deep Learning(DL) approach to improve the accuracy of the results.

1.4 Application:

There are wide range of application areas where this project can be used. It can be used in web mining by extracting only important text from web pages. Traders can use it to track customer reviews. User can check opinion about product by doing Sentiment Analysis(SA) on reviews or text.

exploited by Deep Learning(DL) where upper level concepts that are abstract are being learned from the lower level ones. These architectures are often constructed with a greedy layer-by-layer method. Deep Learning (DL) helps to disentangle these abstractions and pick out which features are useful for learning [15].

2. Deep Learning (DL) & NLP Task

Natural Language Processing (NLP) (NLP) is the most emerging technologies of this era. As Natural languages are ambiguous in nature due to which understanding them is difficult. Natural Language Processing (NLP) is the field of artificial intelligence and Machine Learning(ML). Applications of NLP are everywhere it is used in our everyday life: web search, advertisement, emails, customer service, language translation, text summarization etc. There are a large variety of tasks and Machine Learning (ML) models used by NLP applications. Recently, Deep Learning (DL) approaches are giving very high performance across many different NLP tasks.

Types of Sentiment Analysis (SA):

1. Document level Sentiment Analysis (SA): In document level Sentiment Analysis (SA), opinion is contained from single opinion holder and whole document is focused on single entity.Usually opinions are expressions that are subjective in nature and tell person's review or feeling toward that particular entity.

2. Sentence level Sentiment Analysis(SA): Sentence level sentiment is more refined form of document level analysis. To determine whether the opinion on entity is positive or negative, we filter out those sentences that do not express any sentiment about that entity.

3. Aspect level Sentiment Analysis (SA): Document and sentence level Sentiment Analysis (SA) works well where whole document focuses on single entity. However, if the opinion is about entities considering multiple aspects then we should move to aspect level Sentiment Analysis (SA). For example, "I like Sony phone. Its camera quality is amazing but the battery life sucks" the above sentence has mixed reviews about entity Sony phone. For this, we need to move to aspect level through which can recognize review about different attributes of entity Sony phone. In aspect level Sentiment Analysis (SA) opinion about different aspects in document are recognized.

4. Comparative Sentiment Analysis (SA): In document, some of the sentences contain comparative opinion. Opinion is expressed by comparing products with some other product. For example, "The camera quality of Sony is better than Samsung" this sentence is comparing aspect of different entities. We need to identify such type of sentences in Comparative Sentiment Analysis (SA).

Applications of Sentiment Analysis(SA):

Sentiment Analysis (SA) has broad application area. Mainly it is very useful in monitoring social media reviews and blogs. Some of the Sentiment Analysis (SA) applications are as follows:

1. Focal points are Twitter and Facebook. Sentiment Analysis (SA) can be used to determine the opinion of tweets whether the tweet is positive or negative.

2. Monitoring social media and tracking user reviews

3. Based on news and blogs, it can be used for market movement forecasting in financial domain.

4. Used to compute satisfaction of customers i.e. to identify whether the customers are happy with product or not.

5. To predict trends Sentiment Analysis (SA) can be used.

6. Useful in recommendation systems.

There are mainly three approaches to acquire sentiment lexicon:

1. Manual Approach: In manual approach, people code sentiment lexicon manually. However, this is not feasible approach as very large lexicon is required and much laborious effort is needed to create such lexicon for each domain. This approach is rarely followed.

2. Dictionary-based approach: The dictionary based approach starts with a small set of seed sentiment words suitable for domain [28]. Further, using various features of Word Net or dictionary like synonyms and antonyms these set of words can be expanded. For this, one algorithm is given in Kamp et al [29]. The limitation of this approach is that the lexicon acquired using this approach is not domain specific and hence does not capture specific peculiarities of any specific domain[28].

3. Corpus-based approach: Corpus based approach is followed when one wants to create domain specific lexicon. In this approach a large domain specific corpus is needed to create a lexicon.

Challenges in Sentiment Analysis (SA):

Although Sentiment Analysis(SA) and opinion mining are widely studied application areas of NLP. However, there are various major challenges faced while analyzing opinion about entities. Some of them are as follows:

1. As, words behave differently in different context. It is difficult to identify sentiment-expressing words from text i.e. subjective words are difficult to identify from text. For example:

"His language is very crude."

"Inventories hit a record despite draw downs in crude oil storage."

In above sentences the word crude is opinion-expressing word in first sentence whereas it is objective in second sentence.

2. The major challenge faced in Sentiment Analysis (SA) is words domain dependency. Some sentiment words have different meaning in different domain. It is difficult to identify whether the word is used in positive sense or negative sense. For example the words "unpredictable" if it is used in movie domain it is positive but if word is used in vehicle domain then word acts as negative

3. Sarcasm Detection is very difficult. For example, "Not all men are annoying. Some are dead" identifying polarity of these sentences is difficult.

3.Recent Research in NLP using Deep Learning(DL)

- Facebook launches Advanced AI effort to find meaning in your post[17]
- Google Virtual Brain Goes to work.[17]
- Microsoft brings star treck'svoice translator to life[17]
- Enlitic picks up $2M to help diagnose diseases with Deep Learning(DL)[18]
- Butterfly Network Hopes to Bring Deep Learning(DL) AI to Medicine[18]
- A Googler's Quest to Teach Machines How to Understand Emotions.[18]

This system is provisioned to be built using ANT Tool &Restful Web Services that is highly flexible. In this project user can upload only specific file formats and in Text Summarization file size must be less than 3MB and user input text must be less than 20000 characters.

3.1 Architecture:

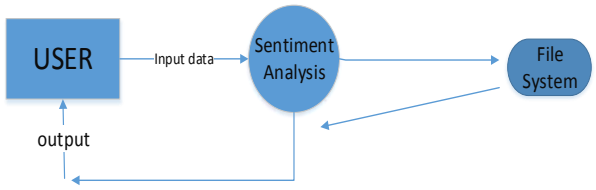

Figure 1 System Architecture

Login module consists of two sub modules sign in and sign up. If the user is already registered then user has to enter Email Id /user Id and password to access his/her account. Else first the user has to create an account by signup.

Flow chart of login module:

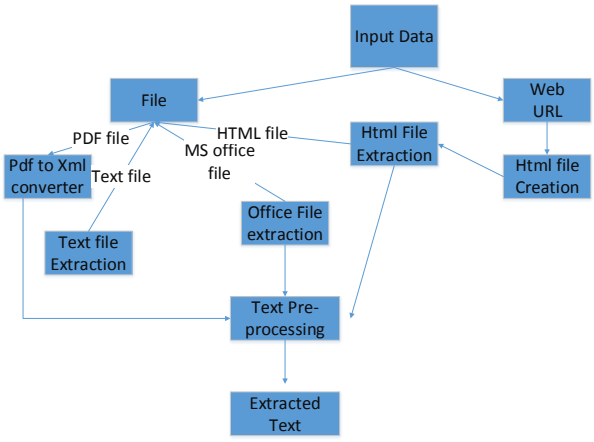

Figure 1 Text Extraction Flow Chart

Procedure followed is as follows:

1. User has both options he/she can upload file of different format (PDF, text, MsOffice or Html) or write a web URL from where he/she wants to extract text.
2. After that, the user input is checked whether valid URL or Valid File format is there or not.
3. If input URL is valid then html file of that web page is created.
4. If file format is valid then we are extracting the file type and language of file.

Steps performed are:

1. Sentence Segmentation phase: In this phase, whole document is divided into sentences and polarity of each sentence is calculated separately.

2. Dependency Parsing phase: After segmentation of all sentences, we are performing dependency parsing of those sentences. These parser connects words according to their relationship.

For example, "This movie is fantastic".

Dependency graph of above sentences is: [det (movie-1, This-1), nsubj(fantastic-4,movie-2),cop(fantastic - 4, is-3), root (Root-0, fantastic-4)]

3. Dependency Graph Feature Score: Once we are done with dependency parsing of a sentence. We are checking some dependency relations on the parse tree to determine the significant relations. Depndency relations are significant if it involves:

i. Any nsubj, nobj and agent relation.

ii. Any adjectival component like acomp.

iii. Any negation relation like neg.

iv. Any modifier or dependent relation like advcl, advmod, amod, dep.

These relations helps to identify which words are closely related to each other. Two words $w_i$and w_j are directly related to each other if there is any dependency relation between them.

The dependency relation nsubj, nobj are used to identify entities of the text i.e. the opinion list which represent text is about what.

4. Rule Feature Score: After calculating, the graph score if it came out to be zero then rule feature score is taken into consideration. To calculate the rule feature score. We are pos tagging the sentences, segmenting sentences into words, and extracting the words score based on their pos tag from SentiWordNet and if the word is not, there in SentiWordNet then word score is extracted from our own lexicon. Then the score of all the words except stop words are added and stored in rule feature score.

5. Phrases feature score: In this, we are handling phrases in sentence. Phrase feature and rule feature scores are calculated simultaneously.

6. Emotion feature score: We are also handling smiley in document. If any smiley is present in sentence then it is separated from sentence and its score or polarity is extracted from emoticon lexicon.

7. Score of each Sentence: After calculating all the above scores final score of each sentence is calculated. It consists of three scores score0,

score1, and score2. These scores are as follows:

score0=Graph_Feature_Score;
if Graph_Feature_Score=0 then

score0=Rule_Feature_Score;
score1=Phrase_Feature_Score;
score2=Emotion_Feature_Score;
Score= score0+score1+score2;

8. Final score: Once we got the Score of each sentences then final score of complete document is computed.

Final_Score= $\sum_1^n scorei$
where 1<=i<=n, n is total number of sentences Steps performed are as follows:

1. Sentence Segmentation phase: In this phase, whole document is divided into sentences and polarity of each sentence is calculated separately.

2. Feature Scores: For Hindi Sentiment Analysis(SA), we are taking four feature scores to calculate the score of each sentence. Feature scores are:

2.1 Rule Feature Score: To calculate the Rule feature score some rule are formed based on pos tag of the words in sentences. Mainly we are focusing on Adjective, Adverb, Verb, and Noun as these are the only words that express opinion in sentence. Rule are formed according to the semantic property of Hindi sentences in which when two same polarity words or opposite polarity words are present in sentence it enhance the polarity of sentence. Rules are as follows:

Table 1 Hindi Sentiment Rules

Relation	**Example**	**Phrase Polarity**
PosAdj + PosAdj	खुबसूरत(+1) मनोरम(+1)	**2**
PosAdj + Neg Adj	अच्छाखासा(+1) नुकसान(-1)	**-2**
PosAdj + Neg Verb	बहुत(+1) भारीपड़ना(-1)	**-2**
Pos Adverb + Neg Adj	पुर्णतः(+1) निरस्त(-1)	**-2**
PosAdj + Neg Noun	ज्यादा(+1) दुःख(-1)	**-2**
Neg Adj + Pos Noun	भयंकर(-1) उत्साह(+1)	**2**
Neg Adj + PosAdj	कम(-1) ज्यादा(+1)	**2**
Neg Adj + Neg Adj	घोर(-1) अपमान(-1)	-2

2.2 Negation Feature Score: Negation feature score is taken to check whether any word that reverses the polarity of sentence is present in sentence or not. In Hindi sentences presence of negative word after any word reverses its polarity i.e. if the polar word is negative negation feature reverses that word to positive and vice versa. We are setting the sentence negation feature score to true if "NEG" tag is present in the pos tag of that sentence. This negation feature score is used further while calculating the final score of sentence.

Negation_Feature_Score = true, if pos tag contains NEG
Negation_Feature_Score = true, otherwise

2.3 Phrase Feature Score: Phrase has their own polarity. Sometimes individual words do not have any polarity but when they are combined together and become phrase they have some polarity

2.4 Emotion Feature Score: We are handling smileys in Hindi document also. If any smiley is present in sentence then it is separated from sentence and its score or polarity is extracted from emoticon lexicon.

3. Score of each Sentence: After calculating all the feature scores. Score of each sentence is calculated. It consists of four scores score0, score1, score2, score3. These scores are as follows:

score0=Rule_Feature_Score;

score1=Phrase_Feature_Score;

score2=Emotion_Feature_Score;

Score= score0+score1+score2;

If Negation_Feature_Score = true, then Score= -Score;

4. Final score: After computing, the score of each sentences individually. Final score of complete document is calculated by adding the scores of all the sentences in document.

Final_Score= ∑_1^n▒scorei

where 1<=i<=n, n is total number of sentences

3. Opinion Detection Phase: Opinion Detection phase is to find the polarity or opinion of complete document whether the document is positive, negative and neutral. The polarity of document is computed based on final score computed in language detection phase.

Polarity = positive, if Final_Score> 0

Polarity = negative, if Final_Score<0

Polarity = neutral, otherwise

Results:Two metrics precision and recall are used to measure the performance of proposed system. Precision measure the exactness of sentiment classification. Recall measures the completeness of classification. Lower precision means less false positive and high recall means less false negatives. One combine metric F-Measure is used by combining Recall and Precision. Recall=(Total number of correctly classified document)/ (Total number of correct document)

Precision = (Total number of correctly classified document)/ (Total number of classified document)

We have evaluated system on 1000 movie reviews and out of them 750 documents are correctly classified i.e. positive document are classified as positive and negative documents are classified as negative.

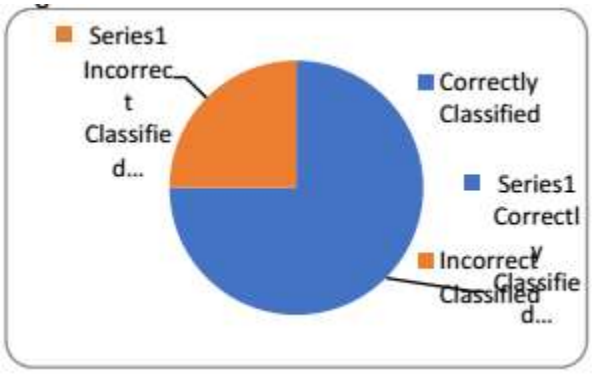

Figure 15 Sentiment Analysis (SA) Test

Conclusion:

In this work, Analysis on text will be done using Deep Learning(DL) approach. Various NLP tasks such as Sentiment Analysis(SA)will be implemented in this Project. Through the text extraction module, main contents are extracted from any web page or file and can be used to perform other operation.

Sentiment Analysis(SA) require dependency parser and lexicons to classify the document as positive, negative or neutral.

Reference

[1] S. P. Yong, A. I. Z. Abidin and Y. Y. Chen, "A Neural Based Text Summarization System", 6th International Conference of Data Mining, pp. 45-50, 2005.

[2] Li Chengcheng, "Automatic Text Summarization Based On Rhetorical Structure Theory", International Conference on Computer Application and System Modeling (ICCASM), vol. 13, pp. 595-598, October 2010.

[3] Naresh Kumar Nagwani, Dr. ShrishVerma,"A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 17– No.2, March 2011.

[4] Hongyan Jing, "Sentence Reduction for Automatic Text Summarization", In Proceedings of the 6th Applied Natural Language Processing(NLP) Conference, Seattle, USA, pp. 310-315, 2000.

[5] Nitin Agarwal, Gvr Kiran, Ravi Shankar Reddy and Carolyn PensteinRos´e, "Towards Multi-Document Summarization of Scientific Articles: Making Interesting Comparisons with SciSumm", Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, Portland, Oregon, pp. 8–15.

[6] Jun'ichi Fukumoto, "Multi-Document Summarization Using Document Set Type Classification", Proceedings of NTCIR- 4, Tokyo, pp. 412-416, 2004.

[7] P. Turney, "Thumbs up or thumbs down? Semantic orientation Applied to Unsupervised Classification of Reviews", In Proceedings of the Association for Computational Linguistics, pp.417-424, Philadelphia, 2002.

[8] E. Cambria, A. Hussain and C. Eckl, " Taking Refuge in Your Personal Sentic Corner", In Proceeding of Workshop on Sentiment Analysis(SA) where AI meets Psychology, pp.35-43, Thailand, 2011.

[9] A. Joshi, A. R. Balamurali and P. Bhattacharyya, "A Fallback Strategy for Sentiment Analysis(SA) in Hindi: a Case Study", In Proceedings of the 8th ICON, 2010.

[10] PadmaPriya, G. and K. Duraiswamy, "An Approach for text summarization using Deep Learning(DL) algorithm", Journal of Computer Science 10 (1): 1-9, 2014

[11] Joel LaroccaNeto Alex A. Freitas Celso A. A. Kaestner,"Automatic Text Summarization using a Machine Learning(ML) Approach".

[12] NeelimaBhatia,ArunimaJaiswal,"Literature Review on Automatic Text Summarization: Single and Multiple Summarizations", International Journal of Computer Applications (0975 – 8887)Volume 117 – No. 6, May 2015.

[13] J.SHARMILA,A.SUBRAMANI, "A Methodfor extracting information from the web using Deep Learning(DL) algorithm", Journal of Theoretical and Applied Information Technology 20th October 2014. Vol. 68 No.2.

[14] RDRPOSTagger , Rule-based Part-of-Speech and Morphological Tagging Toolkit , Dat Quoc Nguyen, Dai Quoc Nguyen, Dang Duc Pham, and Son Bao Pham.

[15] https://en.wikipedia.org/wiki/Deep_Learning, (Accessed: 15 July, 2015).

[16] Collobert R., J. Weston, et al. (2011). "Natural Language Processing(NLP) (almost) from scratch." The Journal of Machine Learning(ML) Research 12: 2493-2537.

[17] XiaodongHe,Jianfeng Gao and Li Deng,"Deep Learning(DL) for Natural Language Processing(NLP) and related application".

[18] http://www.slideshare.net/roelofp/220115dlmeetu, (Accessed: 20 July, 2015).
[19] https://tika.apache.org, (Accessed: 6 August, 2015)
[20] https://word2vec.googlecode.com/svn/trunk, (Accessed: 7 October, 2015).
[21] https://www.gavagai.se/blog/2015/09/30/a-brief-history-of-word-embeddings (Accessed: 23 September,2015).
[22] https://en.wikipedia.org/wiki/Word_embedding, (Accessed: 22 September,2015).
[23] https://opennlp.apache.org, (Accessed: 16 November,2015).
[24] Cacm.acm.org/magazines/2013/4/162501-techniques-and-applications-for-sentiment-analysis/fulltext
[25] Kamps, J. Marx, M. Mokken, R.J. and de Rijke, M. using WordNet to measure semantic orientation of adjectives. LREC,2004.
[26]www.cfilt.iitb.ac.in/resources/surveys/SentimentAnalysis-Vinita.pdf
[27] Pang, B. L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment classification using Machine Learning(ML) techniques", In proceedings of Annual Meeting of the Association for Computational Linguistics, 2009.