IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June Month 2021

DOI 10.17148/IJARCCE.2021.106122

Prediction of Polarity in Online News Articles

Swapna Bhavsar[1], Aditi M Metkar[2], Sourabh Mokashi[3], Kajol Sonawane[4], Akshay Patil[5]

Assistant Professor, Department of Information Technology, PES's Modern College of Engineering Pune 411005, Maharashtra India[1]

Student, Department of Information Technology, PES's Modern College of Engineering Pune 411005, Maharashtra India[2, 3 4, 5]

Abstract: The importance of online news articles has evolved with the advancement of information and technology. How people gather information, shape their views, and engage with topics of relevance has been increased by the internet. Thus news articles become important sources and play a significant role in shaping personal and public opinion. Predicting polarity in news articles becomes crucial to have a well-balanced understanding of any event. Using aspect-based sentiment analysis, the application predicts sentiments attached to various aspects of a particular Hindi news article. Our approach consists of sentence identification using POS tagging techniques followed by aspect extraction using unsupervised learning algorithms and finally predicting the sentiments of the aspects using sentiment analysis. The predicted sentiments would be displayed in a user-friendly format so that the users can easily understand them. Existing systems work on the English language unlike our approach for sentiment analysis

Keywords: News Articles, Polarity, Sentiment Analysis, Topic Modeling, RNN

I. INTRODUCTION

A. News Articles and Polarity

Newspapers have shaped public opinion for the longest of times and have been the major source of information. With the advancement, today we read online news articles and hold our opinions accordingly. Therefore a lot of research has been found for user opinion mining. Several recommendation engines filter out articles to be presented to a user, whereas the reader might be casually looking for a piece of objective information. Drawing out what these articles contain and what sentiment it carries can thus be helpful for each individual.

B. ABSA

Aspect-based sentiment analysis is a technique that considers aspects or attributes of anything that is discussed in a text. The next step is that it analyses the sentiment for each aspect considered i.e. positive, negative, or neutral. Sentiment Analysis focuses on the sentiments of a given text as a whole. This technique has been majorly used over customer reviews of restaurants and movies where they have considered well-defined aspects such as quality of food, ambiance, etc. The sentiments were then determined for concerning aspects. There are four main tasks involved in ABSA as mentioned in [1]: Aspect Term Extraction (ATE), Aspect Term Sentiment (ATS) classification, Aspect Category Detection (ACD) & Aspect Category Sentiment (ACS) classification.

C. Topic Modelling

Topic modeling is a technique by which various topics discussed in a certain text or collection of documents can be extracted. Topics are nothing but a pattern of terms repeating and co-occurring in a corpus. A document contains multiple topics with varied proportions. Applications include document clustering and information retrieval from unstructured text. Organizing large datasets of emails, customer reviews, and social media profiles are some commonly used.

D. Approach/Main Contribution

The main contribution of this paper is to present a multi-model approach for aspect-based sentiment analysis on Hindi news articles. The approach comprises the use of various topic Modeling algorithms.

II. METHODS AND MATERIALS

A. Related Work

There have been various approaches to sentiment analysis as well as for aspect-based sentiment analysis for SEMEVAL dataset of customer reviews of and on IMDB dataset. Majorly these approaches had well-defined aspects.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June Month 2021

DOI 10.17148/IJARCCE.2021.106122

1) Aspect Based Sentiment Analysis in the Hindi Language :

Indian languages especially Hindi has been claimed to be a resource-scarce language due to the non-availability of various tools such as POS tagger, annotated corpora. M.S. Akhtar et.al proposed approaches to reduce the effect of data sparsity[14]. Methodologies for Aspect Category Detection with the multi-label classification framework and Sentiment Classification has been discussed on a dataset created of user web reviews in [13]. They have collected review sentences from various online sources and annotated 5,417 review sentences across 12 domains.

2) Topic Modeling Approaches :

Topic modeling has been used for various goals by researchers such as recommendation systems, emotion classification, image classification, event detection, etc. There have been approaches for aspect extraction methods using various topic modeling algorithms such as LDA, LSA, etc. P. P. Patil et.al (2019) have used unsupervised machine learning methods and a frequency-based approach model for topic extraction (aspect detection) and sentiment detection of user reviews[17]. B. Ozyurt et.al presented another novel method known as SS-LDA(Sentence Segment LDA) which is an adaptation of the LDA algorithm. A sentiment dictionary as a resource is required for this method. The dataset used is of the Turkish language[4]. S. J. Das et.al have emphasized effective summarization of reviews in a structured manner using aspect-based sentiment analysis with a focus on aspect quality. Aspect extraction is done using LDA and the result is being compared with the earlier word2vec method [18]. M. Shams et.al introduced a combination of LDA with word co-occurrence analysis called Enriched LDA for the aspect extraction. The proposed ELDA model is said to be language independent and incorporates the advantages of both methods used in combination. The model also involves relevant topics to enhance knowledge extraction in a small corpus [7].LDA has been used predominantly for topic modeling. Y. Kalepalli et.al have proposed a comparison of LSA and LDA where both the algorithms are applied on the BBC news dataset. Firstly, the documents are preprocessed and converted into a bag of words and then fed to the algorithm. The accuracy level for LDA was found to be 82.57 which is far better than LSA at 75.30 [21]. Similarly, a comparison of NMF and LDA for topic modeling has been analyzed for the covid-19 corpus by S. Mifrah et.al. In terms of coherence of topics generated by both these algorithms from the corpus, LDA proved to be better than NMF. Also, the words for each topic for the LDA model were more significant than the NMF model[10].

3) Comparison of Aspect Extraction Methods :

TABLE I Unsupervised methods

Model	Dataset	Algorithm used
ATE-SPD(aspect-terms extraction and sentiment polarity detection)[1]	Restaurant and laptop reviews dataset from SEMEVAL'14	Bi-LSTM hybridized with CRF
A hybrid unsupervised model with an attention mechanism.[12]	SemEval-16 dataset	Two-step model: Firstly rule-based methods for ATE, then they are used for training attention-based models.
Hybrid unsupervised model for ATE and OTE[6]	SemEval-2014	a chunk-level extraction method for ATE, a domain correlation measurement
Feature extractor and classifier for unsupervised ATE [33]	SemEval-2014	B-LSTM & CRF using Automatically Labelled Datasets
LDA-CRF model[2]	Hotel and restaurant reviews from TripAdvisor.com.	A supervised sequence modeling component via CRF for aspect– sentiment phrase extraction
SVM and ME Model[15]	Laptop review dataset from Amazon.com	SVM with selected lexical and semantic features
CNN model[16]	Laptop review dataset in SemEval-2014 ,restaurant review dataset in SemEval-2016.	CNN with two-word embedding layers, one is general and the other domain-specific.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June Month 2021

DOI 10.17148/IJARCCE.2021.106122

4) Sentiment Analysis Approaches:

For sentiment analysis, there have been various approaches such as lexicon-based and machine learning methods. S. Sharma et.al have performed sentiment analysis on Hindi news articles and reviews using machine learning techniques. Polarity detection of reviews is performed using POS tagging [11].

S. Rani et.al have discussed a deep learning-based sentiment analysis model using CNN. The model consists of four layers i.e. input layer, convolution layer, global max pool layer, and fully connected layer. The results suggest that if the model is properly trained it can do much better than the baseline machine learning model [9] .

A lexicon-based approach for sentiment analysis on news articles has been put forth by S. Taj et.al . The sentiments which are conveyed by each word are deduced by the underlying polarity and the polarity of a sentence is the total sum of polarities of each word in it. For calculating polarity of text, after preprocessing, a technique known as Term Frequency -Inverse Document Frequency(TF-IDF) is used in which the most frequently occurring words are assigned a weightage. Each such word is assigned a sentiment score using a Wordnet dictionary [19].

TABLE II Sentiment Analysis Models

Model	Dataset	Algorithm used	Results
SVM model[5]	Newspaper headlines	Three approaches-Linear SVM, TF-IDF and Linear SVM, SGD classifier	Tf-idf and Linear SVM provides better accuracy for smaller dataset and SGD and linear SVM model for larger dataset
NB model[8]	Amazon product review dataset	Naive Bayes	The approach provides an accuracy of 85.7%
CNN model[9]	Hindi movie reviews from online newspapers and Websites.	Convolutional Neural Network with different combinations of convolution and hidden layers.	CNN model having 2 convolutional layers with filter sizes 3 and 4 performs the best with an accuracy of 95%.
Semantic Orientation model[20]	Book review datasets from GoodReads and Amazon	(Semantic Orientation - Pointwise Mutual Information - Information Retrieval)	The approach gives better accuracy for the Goodreads dataset than Amazon.

B. Workflow

The workflow of our proposed system is as shown in fig1 which consists of various stages such as Translation, Preprocessing, POS tagging, ATE, and SA.

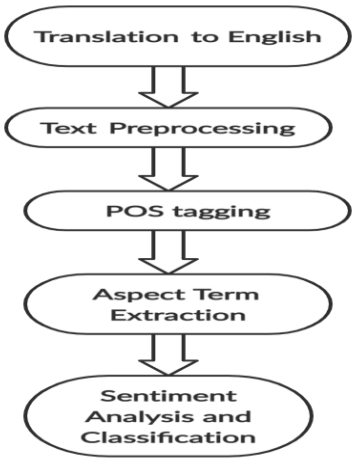

Fig 1. Outline of workflow

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June Month 2021

DOI 10.17148/IJARCCE.2021.106122

1) Translation and Text Preprocessing :
The Hindi news dataset was translated to English using Google Cloud Translate API. The translated text would undergo preprocessing. Preprocessing of text is essential in order to clean the text and use it for the aspect-based sentiment analysis task further. It would involve various steps such as Tokenization, Lemmatization, Stop words Removal, and punctuation removal.

2) POS Tagging :
Part-of-speech-tagging is a mechanism where the words in the given text are tagged grammatically. The words are labeled as a noun, verb, adjective, adverb, etc. It does fine-grained tagging like 'noun-plural" and considers tense while tagging. It plays an important role in finding the aspect words.

3) Aspect Extraction via Topic modeling :
The task of aspect extraction is performed using a multi-model approach which consists of topic modeling algorithms. In this approach, three algorithms will be used for pulling out various topics of a given text which are LDA(Latent Dirichlet Allocation), LSA(Latent Semantic Analysis), and HDP(Hierarchical Dirichlet Process). The topics extracted from them will be combined to determine the aspects of the given text.

4) Sentiment Analysis :
Sentiment Analysis is performed for each extracted aspect using an NLTK sentiment intensity analyzer. It works using VADER, which is a list of words that have a sentiment associated with each of them.

III. Results and Discussion

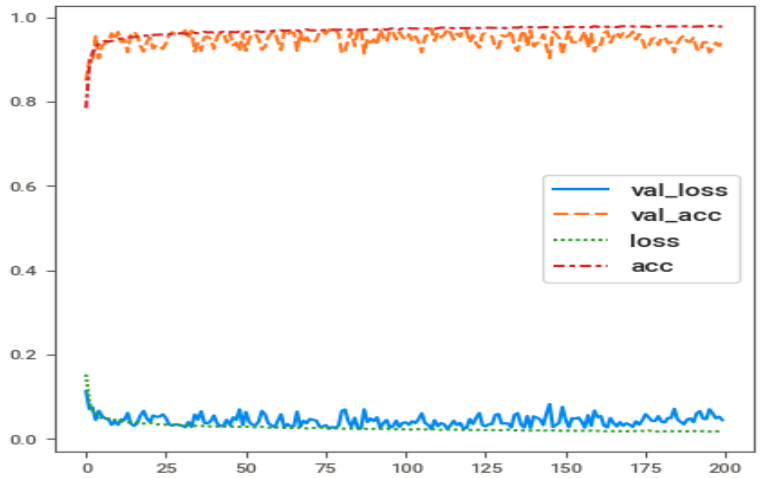

Fig. 2 Accuracy graph

We have performed three different ways for the whole process and summarised them as follows:
1. In the first approach, we preprocessed the Hindi language text using the existing tools and libraries and then performed aspect extraction using topic modeling. The results were not satisfactory since Hindi being a resource-scarce language the existing preprocessing tools do not give expected results.
2. We used the TF-IDF factor to recognize the relevance of a word in a given text on the same data in the previous approach. The mixture of topic modeling algorithms such as LSA(Latent Semantic Analysis), LDA(Latent Dirichlet allocation), and HDP(Hierarchical Dirichlet process) was applied to the data and the results were a bit satisfactory.
3. In this approach, we eliminated the use of Hindi preprocessing tools by translating the text into the English language. The translation accuracy increased when the text was divided into single sentences for translation. We then preprocessed the text and applied the topic modeling approaches from the previous approach. The results improved significantly.
4. We have obtained an accuracy of 97.83% for our final approach.

IV. Conclusion

In this paper, we proposed two different tasks: first aspect term extraction and second finding the sentiment of various aspects involved in a given Hindi text. For aspect term extraction we have used the Multi-modelling approach of topic modeling algorithms in order to overcome the limitation of a single algorithm and draw out the implicit aspects which

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June Month 2021

DOI 10.17148/IJARCCE.2021.106122

are hidden in the text. To predict the sentiment for each of the identified aspect terms, we have used the NLTK sentiment analyzer which provides good accuracy. We have also analyzed three different approaches for the task.

REFERENCES

[1] A. Kumar, S. Verma, and A. Sharan, "ATE-SPD: simultaneous extraction of aspect-term and aspect sentiment polarity using Bi-LSTM-CRF neural network," J. Exp. Theor. Artif. Intell., vol. 33, no. 3, pp. 487–508, 2021, doi: 10.1080/0952813X.2020.1764632.
[2] A. Laddha and A. Mukherjee, "Aspect opinion expression and rating prediction via LDA-CRF hybrid," Nat. Lang. Eng., vol. 24, no. 4, pp. 611–639, 2018, doi: 10.1017/S135132491800013X.
[3] A. Parlina, K. Ramli, and H. Murfi, "Exposing emerging trends in smart sustainable city research using deep autoencoders-based fuzzy c-means," Sustain., vol. 13, no. 5, pp. 1–28, 2021, doi: 10.3390/su13052876.
[4] B. Ozyurt and M. A. Akcayol, "A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA," Expert Syst. Appl., vol. 168, no. March, p. 114231, 2021, doi: 10.1016/j.eswa.2020.114231.
[5] C. J. Rameshbhai and J. Paulose, "Opinion mining on newspaper headlines using SVM and NLP," Int. J. Electr. Comput. Eng., vol. 9, no. 3, pp. 2152–2163, 2019, doi: 10.11591/ijece.v9i3.pp2152-2163.
[6] C. Wu, F. Wu, S. Wu, Z. Yuan, and Y. Huang, "A hybrid unsupervised method for aspect term and opinion target extraction," Knowledge-Based Syst., vol. 148, pp. 66–73, 2018, doi: 10.1016/j.knosys.2018.01.019.
[7] M. Shams and A. Baraani-Dastjerdi, "Enriched LDA (ELDA): Combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction," Expert Syst. Appl., vol. 80, pp. 136–146, 2017, doi: 10.1016/j.eswa.2017.02.038.
[8] R. Vijay, B. Vangara, K. Thirupathur, and S. P. Vangara, "Opinion Mining Classification using Naive Bayes Algorithm," Int. J. Innov. Technol. Explor. Eng., vol. 9, no. 5, pp. 495–498, 2020, doi: 10.35940/ijitee.e2402.039520
[9] S. Rani and P. Kumar, "Deep Learning Based Sentiment Analysis Using Convolution Neural Network," Arab. J. Sci. Eng., vol. 44, no. 4, pp. 3305–3314, 2019, doi: 10.1007/s13369-018-3500-z.
[10] S. Mifrah and E. H. Benlahmar, "Topic modeling coherence: a comparative study between lda and nmf models using covid'19 corpus," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 4, pp. 5756–5761, 2020, doi: 10.30534/ijatcse/2020/231942020.
[11] S. Sharma, S. K. Bharti, and R. K. Goel, "Sentiment Analysis of Indian Language," Int. Res. J. Eng. Technol., pp. 4251–4253, 2018, [Online]. Available: www.irjet.net.
[12] G. Singh Chauhan, Y. Kumar Meena, D. Gopalani, and R. Nahta, "A two-step hybrid unsupervised model with attention mechanism for aspect extraction," Expert Syst. Appl., vol. 161, p. 113673, 2020, doi: 10.1016/j.eswa.2020.113673.
[13] M. S. Akhtar, A. Ekbal, and P. Bhattacharyya, "Aspect based sentiment analysis: category detection and sentiment classification for hindi," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9624 LNCS, pp. 246–257, 2018, doi: 10.1007/978-3-319-75487-1_19.
[14] M. S. Akhtar, P. Sawant, S. Sen, A. Ekbal, and P. Bhattacharyya, "Solving data sparsity for aspect based sentiment analysis using cross-linguality and multi-linguality," NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 572–582, 2018, doi: 10.18653/v1/n18-1053.
[15] N. U. Pannala, C. P. Nawarathna, J. T. K. Jayakody, L. Rupasinghe, and K. Krishnadeva, "Supervised learning based approach to aspect based sentiment analysis," Proc. - 2016 16th IEEE Int. Conf. Comput. Inf. Technol. CIT 2016, 2016 6th Int. Symp. Cloud Serv. Comput. IEEE SC2 2016 2016 Int. Symp. Secur. Priv. Soc. Netwo, pp. 662–666, 2017, doi: 10.1109/CIT.2016.107.
[16] P. Barnaghi, G. Kontonatsios, N. Bessis, and Y. Korkontzelos, Aspect Extraction from Reviews Using Convolutional Neural Networks and Embeddings, vol. 11608 LNCS. Springer International Publishing, 2019.
[17] P. P. Patil, S. Phansalkar, and V. V. Kryssanov, Topic modelling for aspect-level sentiment analysis, vol. 828. Springer Singapore, 2019.
[18] S. J. Das and B. Chakraborty, "An Approach for Automatic Aspect Extraction by Latent Dirichlet Allocation," 2019 IEEE 10th Int. Conf. Aware. Sci. Technol. iCAST 2019 - Proc., pp. 1–6, 2019, doi: 10.1109/ICAwST.2019.8923417.
[19] S. Taj, B. B. Shaikh, and A. Fatemah Meghji, "Sentiment analysis of news articles: A lexicon based approach," 2019 2nd Int. Conf. Comput. Math. Eng. Technol. iCoMET 2019, pp. 1–5, 2019, doi: 10.1109/ICOMET.2019.8673428.
[20] V. D. Kaur, "Sentimental Analysis of Book Reviews using Unsupervised Semantic Orientation and Supervised Machine Learning Approaches," Proc. 2nd Int. Conf. Green Comput. Internet Things, ICGCIoT 2018, pp. 519–524, 2018, doi: 10.1109/ICGCIoT.2018.8753089.
[21] Y. Kalepalli, S. Tasneem, P. D. P. Teja, and S. Manne, "Effective Comparison of LDA with LSA for Topic Modelling," Proc. Int. Conf. Intell. Comput. Control Syst. ICICCS 2020, no. Iciccs, pp. 1245–1250, 2020, doi: 10.1109/ICICCS48265.2020.9120888.
[22] Amazon book review dataset and goodreads https://www.kaggle.com/gnanesh/goodreads-book-review ,Amazon Book Reviews (WebScraped) | Kaggle/
[23] Amazon review dataset http://snap.stanford.edu/data/web-Amazon.html/
[24] BBC hindi news dataset Insight - BBC Datasets (ucd.ie) META-SHARE (ilsp.gr)
[25] Hindi movie reviews from online newspapers and Websites.http://aajtak.intoday.in/film-review.html,
[26] Hotel and restaurant reviews from TripAdvisor.com Supplementary.zip - Google Drive
[27] Laptop review dataset http://jmcauley.ucsd.edu/data/amazon/
[28] Newspaper headlines The Indian Express - Interstitial http://www.jagran.com/entertainment/reviews-news-hindi.html
[29] Semeval-14 dataset SemEval-2014 ABSA Laptop Reviews - Train Data – META-SHARE (ilsp.gr)
[30] Semeval-16 dataset SemEval-2016 ABSA Restaurant Reviews-English: Test Data-GOLD (Subtask 1) –
[31] User reviews dataset https://www.cicling.org/2016/data/170
[32] User reviews dataset in turkish language Türkiye'nin En Büyük Online Alışveriş Sitesi Hepsiburada.com
[33] A. Giannakopoulos, C. Musat, A. Hossmann, and M. Baeriswyl, "Unsupervised Aspect Term Extraction with B-LSTM and CRF using Automatically Labelled Datasets," pp. 180–188, 2018, doi: 10.18653/v1/w17-5224.

