Materials Today: Proceedings xxx (xxxx) xxx

Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Sentiment analysis, opinion mining and topic modelling of epics and novels using machine learning techniques

Krishna Raj P M, Jagadeesh Sai D

Department of Information Science & Engineering, Ramaiah Institute of Technology, Bengaluru, India

ARTICLE INFO

Article history:
Received 2 April 2021
Received in revised form 26 May 2021
Accepted 1 June 2021
Available online xxxx

Keywords:
Natural language processing
Sentiment analysis
Latent Dirichlet Allocation Social Network Analysis
Vader Sentiment Analysis

ABSTRACT

Sentiment analysis systems can collect and automatically structure unstructured information by collecting public views of services, products, policy, brands, etc. This information is of major value in the fields of marketing analysis, public relations, product reviews, net promoter evaluations, customer feedback and client reviews. Literary works, on the other hand, are less susceptible to computational analysis because there are no immediate commercial incentives. However, similar techniques can be used to evaluate literary work, comprehend the underlying social network, and obtain or validate literary work. This project is about analyzing the book's characters and predicting their characteristics and relationships with one another. A lot of human effort is expended during the adaptation of a novel/book in any form, which is inconvenient and undesirable. Furthermore, the human brain has a tendency to overlook a number of minor details about the events/characters in the book. The scenario described above can frequently result in inaccuracies in the adaptation's plot. As a result, the project is an innovation that aims to aid in the easy and accurate adaptation of a book, making the process much simpler and precise. Machine learning (supervised and unattended) and lexical approaches include the current sentiment analysis techniques. The model's goal is to scan the massive amounts of text in the book. Following digitization, the model will display interesting ideas derived from the given book using a combination of natural language processing, feelings and emotions analysis, and social network analysis methodology.

Selection and peer-review under responsibility of the scientific committee of the 1st International Conference on Computations in Materials and Applied Engineering – 2021.

1. Introduction

By the past and recent literary genres, novels and epics have considered most of the work in the Digital Humanity community as the scope is usually large in terms of time, characters to facilitate and number of events computational analysis Fig. 1 Fig. 2 Fig. 3 Fig. 4 Fig. 5 Fig. 6 Fig. 7.

The computational analysis of great texts can bring to light many interesting patterns, the dynamics of many characters and all the details that man reads. The purpose of the project is to produce efficient visualizations from the data book. These visualisations help us to bring a different perspective to the book. The sentiment flow across the chapters are graphed, and the emotional quotient of various books comes out through the visualization performed.

Human readers who read books for the interesting stories or spiritual messages they contain may be unaware of the overwhelming amount of knowledge found within them. Computational analysis of large texts can reveal fascinating trends and insights into the structure, flow, and dynamics of the many characters in complex stories. The project entails investigating the book's characters and events, as well as anticipating their characteristics, relationships, and connections.

The project aims on generating a generalized model, which it produces effective visualization for any given book. The book needs to be in UTF-8 format. Since the given book is often unstructured, preprocessing techniques like tokenization, stop word removal, lemmatization etc. is done. The structured data is subjected to POS tagging, and further topic modelling using Latent Dirichlet Allocation Algorithm. The project generates a list of all of all characters present in the book. The finely structured data is used for word cloud generation. Given any character from the book, the character's qualities are displayed. Sentimental analysis is performed on the data, using VADER. This generates effective visualization of data in the book.

E-mail addresses: krishnaraj@msrit.edu (K. Raj P M), djsai@msrit.edu (J. Sai D)

https://doi.org/10.1016/j.matpr.2021.06.001
2214-7853/
Selection and peer-review under responsibility of the scientific committee of the 1st International Conference on Computations in Materials and Applied Engineering – 2021.

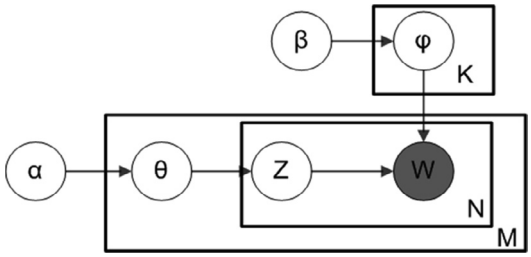

Fig. 1. Plate diagram of LDA.

1.1. Literature review

The different approaches to sentiment analysis are widely spread among two main groups (SA). The first group solves SA problems through the implementation of a machine learning approach. This group uses a number of techniques to extract excellent features that better inform people about the polarity of their feelings. The technique used is monitored continuously because the corpus is manually noted. The second group uses a linguistically inclined method called lexicon-based method. The investigation is initiated according to references with words or phrases that show semantical polarity characteristics. There is also another group that combines master learning with the Lexicon group. The group was referred to as a combination method or semiconductor.

In the paper titled "Improving Part-of-Speech Tagging for NLP Pipelines" [1], the results of sentence-level linguistics-based rules for enhancing part-of-speech tagging were presented by the authors. It was well understood that if one of the preliminary stages was not up to par, the performance of complex NLP systems would suffer. The research gap was discovered to be that current POS taggers work poorly at the sentence stage, and that high token level metrics are misleading.

The author proposed a novel method for segmenting text into tokens and sentences called 'waste' in the paper entitled 'Word and Sentence Tokenisation with Hidden Markov Models' [2]. Hidden Markov Model was approached for the detection of segment borders. The unmodified waste detector did not function properly and showed a lack of research.

In the paper titled "Adapting computation text analysis to social science and vice versa" [5], the author discovered that minor gaps in viewpoint split Social scientists and computer scientists, rather than any major disciplinary divide. Several such variations have been noted in the field of text analysis. The research gap was identified as the need for better ways to choose among possible textual corpus models [6].

In the paper titled "LDA based topic modelling of journal abstracts" [3], the authors discovered that topic modelling was an efficient technique for analysing large document collections without supervision. LDA (Latent Dirichlet Allocation) with Collapsed Variational Bayes and Gibbs sampling were used to create the text models. The designed model was then used to extract relevant abstract tags. The paper's findings revealed that the topics identified by the phrases were simpler to understand than their LDA. However, phrases can contribute low-level events in records, which are impossible to retrieve efficiently.

In the paper titled "Topic Detection from Microblogs Using T-LDA and Perplexity" [4], the Term Frequency-Inverse Document Frequency (TF-IDF) was found by authors and they brought in were able to change the weight of words and measure at a high rate without taking into account the impact of word positions in papers in documents, to assist in the extraction of main words from a short article. The paper developed T-LDA, a new topic detection method that combines LDA and TF-IDF. Experiments on microblog data have demonstrated the approach's efficacy and efficiency in depicting subject evolution. However, the algorithm's performance was not up to mark.

Naïve Bayes and SVM algorithms have also been suggested as classificators. The accuracy gap between positive and negative is closed with unigrams and bigrams. The researchers found that combining machine learning methods and dictionary methods improves the classification of feelings substantially. Sentiment analysis (SSA) always begins with the collection of social media websites like Twitter and the use for sentiment analysis of previous resources such as publicly available data.

2. Research methods

2.1. Data collection

The first and the foremost step is the collection of data. The data in the picture is a story book, which has been encoded using UTF-8 format. For the specified project the English version of Mahabharata was taken from "https://www.sacred-texts.com/hin/maha/index.htm", and English translation of the epic by Mr. KM Ganguly. This particular translation was chosen as it was the only one to cover the entire epic and is considered as its most reliable transla-

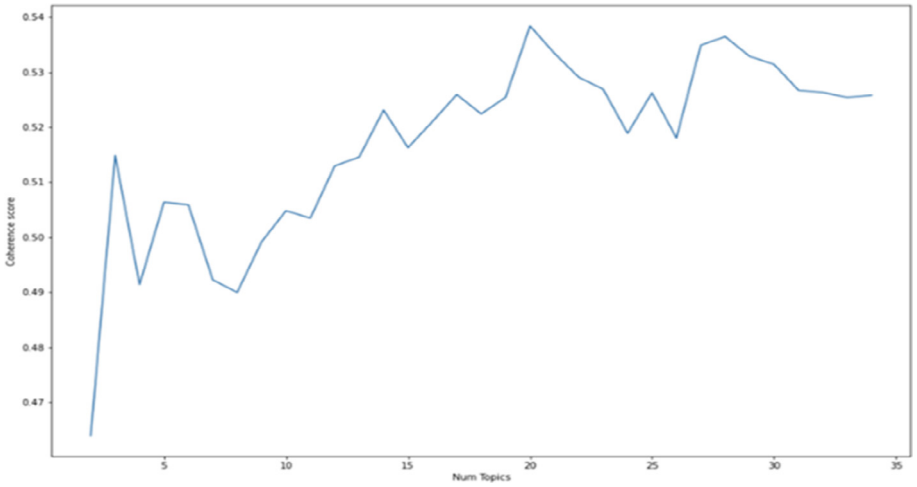

Fig. 2. Elbow plot.

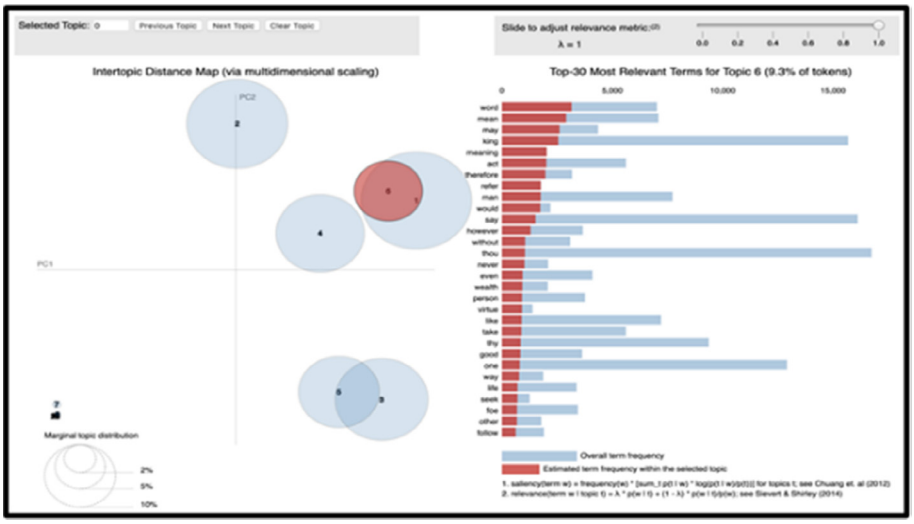

Fig. 3. Topic Modelling Interactive Output Screenshot.

	Dominant_Topic	Topic_Keywords	Num_Documents	Percent_Documents
0.0	0.0	brahmana, man, sacrifice, gift, food, person, merit, give, brahmanas, perform	176	0.0835
1.0	1.0	hath, word, ye, grief, -PRON-, heart, lady, time, address, behold	162	0.0768
2.0	2.0	thou, thee, thy, art, hast, word, address, thine, thyself, dost	77	0.0365
3.0	3.0	son, great, rishi, brahmana, bear, father, wife, ascetic, hear, child	213	0.1010
4.0	4.0	man, act, person, good, virtue, wealth, world, life, foe, duty	211	0.1000
5.0	5.0	god, great, lord, call, foremost, celestial, indra, energy, universe, world	173	0.0820
6.0	6.0	tree, mountain, beautiful, body, head, begin, forest, sun, adorn, elephant	124	0.0588
7.0	7.0	king, yudhishthira, bharata, monarch, great, man, earth, foremost, kingdom, make	146	0.0692
8.0	8.0	soul, mind, knowledge, body, creature, act, sense, form, object, attain	178	0.0844
9.0	9.0	car, son, battle, warrior, arrow, great, king, drona, pierce, steed	359	0.1702
10.0	10.0	son, battle, arjuna, krishna, bhishma, duryodhana, slain, pandu, hero, foe	279	0.1323
11.0	11.0	word, verse, sense, explain, act, call, make, commentator, sacrifice, render	11	0.0052

Fig. 4. Topic Modelling Output for Mahabharata.

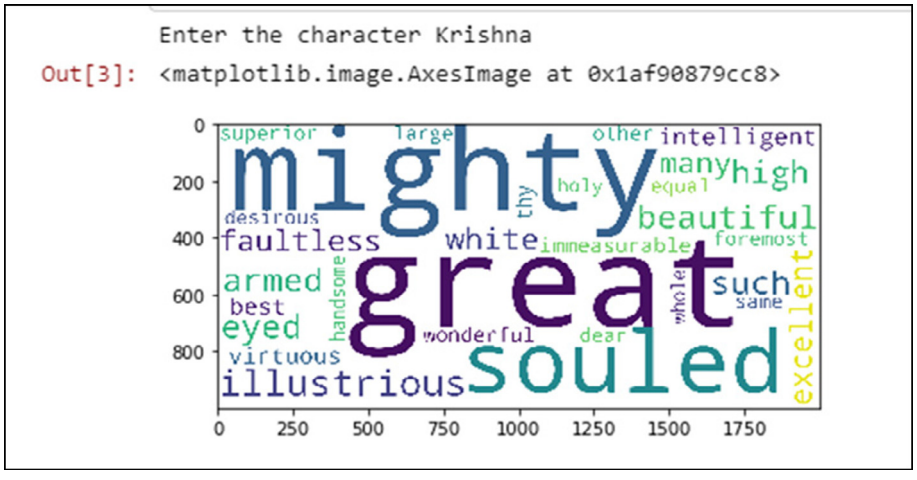

Fig. 5. Word Cloud for Lord Krishna in Mahabharata.

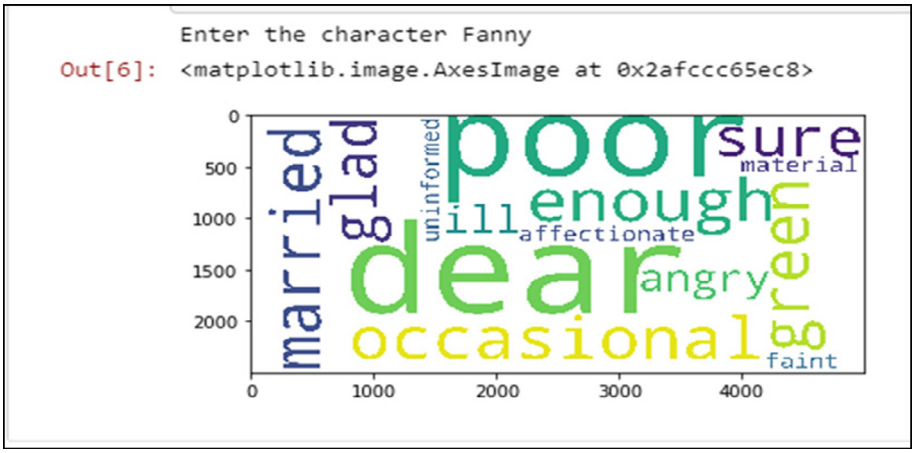

Fig. 6. Word Cloud for Fanny in Sense and Sensibility.

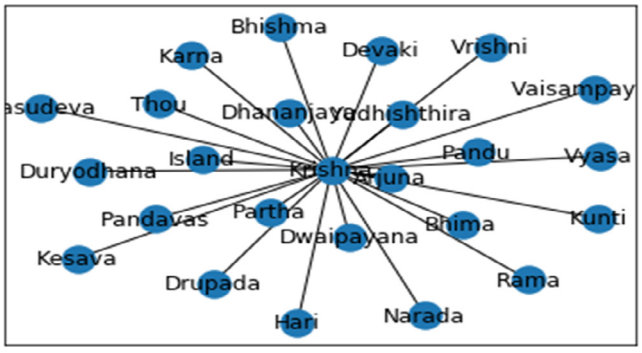

Fig. 7. Network Relationship graph for Lord Krishna Mahabharata.

tion. For getting the UTF-8 encoded versions of English Novels, PROJECT GUTENBERG was used.

2.2. Algorithm: LDA (Latent Dirichlet Allocation)

1. Suppose that all papers have k themes in total.
2. Assign a topic to each word and share these k topics via document m (this distribution is known as α and may be symmetrical or asymmetric, later).
3. Assume that the topic of each word w in document m is incorrect, but that the topic of every other word is right.
4. Assign word w to a subject using probabilistic methods based on two factors:
- which topics are in document m
- across all of the documents how many times word w has been assigned a particular topic (this distribution is called β).
5. This process is kept on repeating number of times for each document and you're done!

2.3. Implementation

The ldamodel by Gensim is the first model used, the coherence score for Gensim was 0.48. This isn't great; in fact, the next algorithm, Mallet, almost always outperforms Gensim's. However, the pyLDAvis, an interactive map that runs in a Jupyter notebook, is a very cool feature of Gensim. It displays the proportion of words in each cluster by plotting the clusters with two principal components.

Mallet (Machine Learning for Language Toolkit), a Java-based package developed by UMASS Amherst, was the next implementation examined. The mallet model for the Mahabharata's 18 books was finally completed. The top ten keywords produced by the model for and latent topic are listed below.

2.4. Vader sentiment analysis

VADER (Valence Conscious Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis platform that is tuned in to social media sentiments. VADER is a form of sentiment analysis that uses lexicons of sentiment-related terms as its foundation. In this method, each word in the lexicon is rated to determine whether it is positive or negative, as well as how positive or negative it is in certain situations. An excerpt from VADER's lexicon as shown below, with more positive terms receiving higher positive ratings and more negative words receiving lower negative ratings.

3. Results and discussion

The post analysis results of the book were tested based on personal knowledge of the book and other related documents obtained from the internet.

The obtained model was tested for: -

(1) Mahabharata

(2) Sense and Sensibility

4. Sentiment analysis graphs

4.1. Detailed sentiment analysis

Fig. 8 Fig. 9 Fig. 10.

4.1.1. Topic modelling & processing time analysis

Fig. 11 Fig. 12 Fig. 13.

4.1.2. File size vs time taken

Fig. 14 Fig. 15 Fig. 16.

Fig. 8. Chapter wise sentiment for Mahabharata.

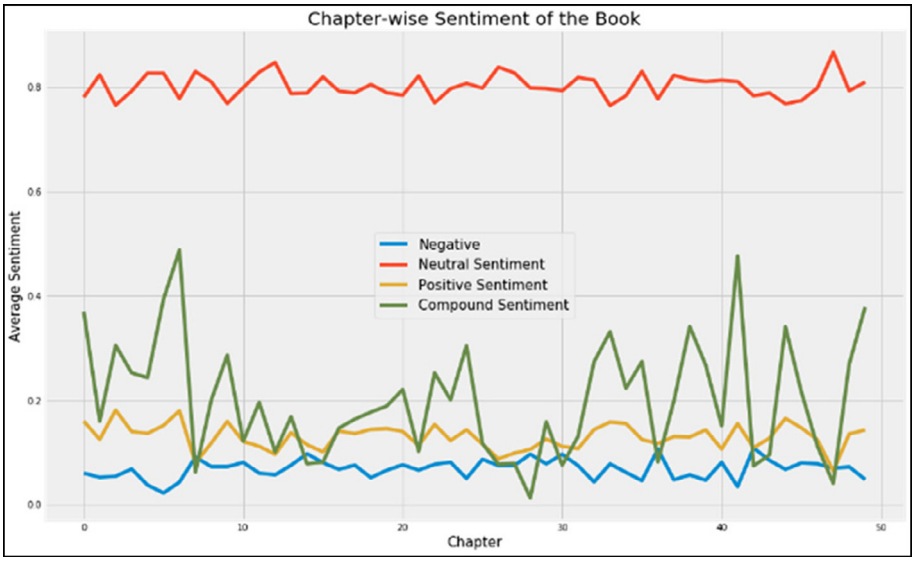

Fig. 9. Chapter wise sentiment for Sense and Sensibility.

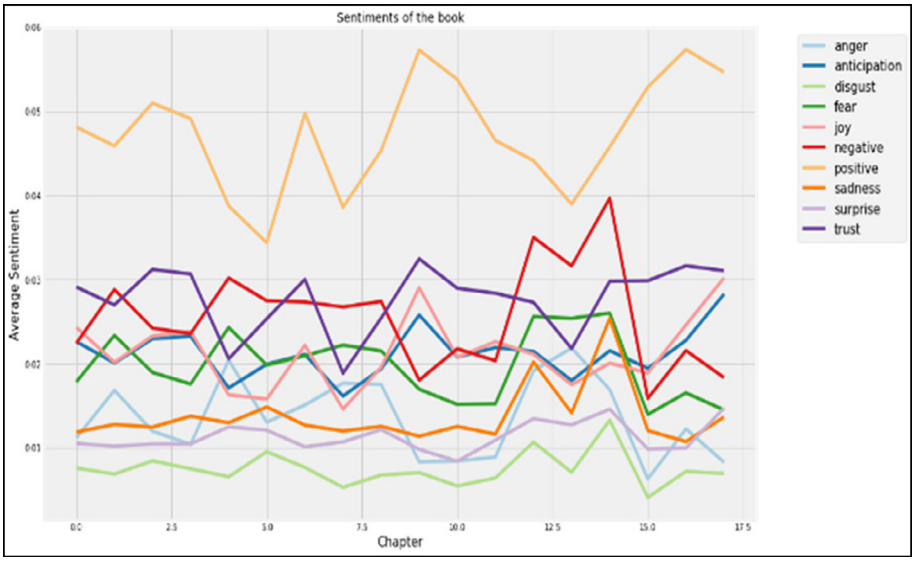

Fig. 10. Detailed Sentiment Analysis for Mahabharata.

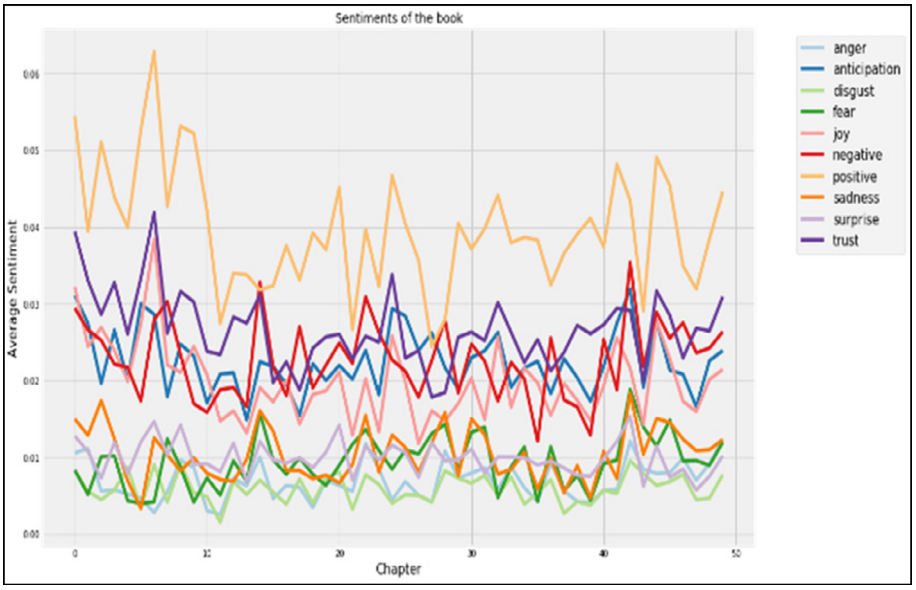

Fig. 11. Detailed Sentiment Analysis for Sense & Sensibility.

	Dominant_Topic	Topic_Keywords	Num_Documents	Percent_Documents
0.0	0.0	brahmana, man, sacrifice, gift, food, person, merit, give, brahmanas, perform	176	0.0835
1.0	1.0	hath, word, ye, grief, -PRON-, heart, lady, time, address, behold	162	0.0768
2.0	2.0	thou, thee, thy, art, hast, word, address, thine, thyself, dost	77	0.0365
3.0	3.0	son, great, rishi, brahmana, bear, father, wife, ascetic, hear, child	213	0.1010
4.0	4.0	man, act, person, good, virtue, wealth, world, life, foe, duty	211	0.1000
5.0	5.0	god, great, lord, call, foremost, celestial, indra, energy, universe, world	173	0.0820
6.0	6.0	tree, mountain, beautiful, body, head, begin, forest, sun, adorn, elephant	124	0.0588
7.0	7.0	king, yudhishthira, bharata, monarch, great, man, earth, foremost, kingdom, make	146	0.0692
8.0	8.0	soul, mind, knowledge, body, creature, act, sense, form, object, attain	178	0.0844
9.0	9.0	car, son, battle, warrior, arrow, great, king, drona, pierce, steed	359	0.1702
10.0	10.0	son, battle, arjuna, krishna, bhishma, duryodhana, slain, pandu, hero, foe	279	0.1323
11.0	11.0	word, verse, sense, explain, act, call, make, commentator, sacrifice, render	11	0.0052

Fig. 12. Topic modelling table for Mahabharata.

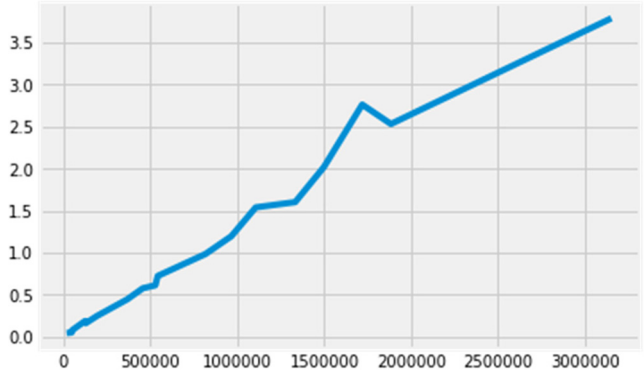

Fig. 13. Number of words vs time taken to process.

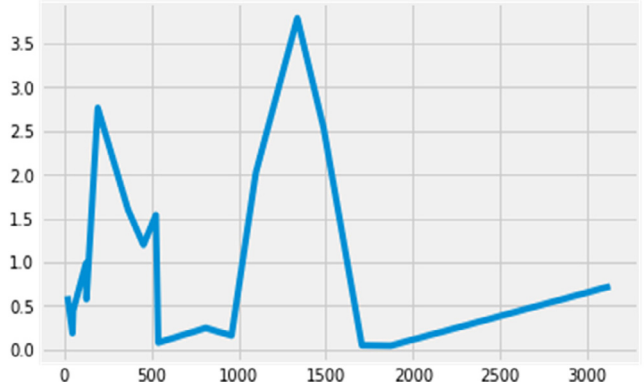

Fig. 14. File size (in kb) vs time taken to process.

4.1.3. Comparing NLP-sentiment analysis with VADER-sentiment analysis

The vader social sentiment analysis algorithm performs 15% better in overall accuracy. The work which is presented in this article is validated using various test processes [7 8] and test techniques [9 10].

5. Conclusion

To put it in simpler words, the project starts with the textual representation of the book and ends with informative visuals. The input to the algorithm is expected to be a digital version of the book. The output of the same is expected to be abundant visualizations portraying different scenarios in the book. These visualizations give us a much clearer idea of underlying events mentioned in the book. With visualizations such as word-cloud and network graphs, a digital and more precise interpretation of a character and its associations with other protagonists can be made. Sentimental Analysis of the same gives us concise information of the sentiments involved in particular actions/events which can give us a much clearer picture of what is happening and why it is happening. The model obtained in the project is very generic and similar visualizations can be obtained for any digital version of the book/document. The model uses LDA as one of the pre-processing techniques which reduces the risks of missing out on an important topic/event while reducing the computation factor at the same time. The main challenges include classification difficulties, widespread speech and negative approaches.

The future work is to further upgrade the model to minimize the computations involved in cases of books having large numbers of sections through pipelining and cloud computations. A Character-generating algorithm can be developed which can be used to generate an image of the character from the adjectives and other POSs used in the book. Full book summarization techniques can be developed which can be used to get a brief summary of the entire book. With NLP and Digital Image Processing techniques combined, one can come up with an algorithm that can be used to convert an entire book into a series of demonstrative and informative videos, which will help portray the theme and mood of the book in a much appealing way. The sentiment analyser is dependent on the language among the approaches examined. No existing method has been found to be more general and more language dependent. However, there is a growing concern in non-English languages with regard to sentimental analysis or opinion mining because research and resources in other languages remain lacking.

Social media sources like microblogs, forums, news and blogs provide plenty of information about people's thoughts and feelings about a particular issue or product. Forum or message board users have to register to send publication documents to a site before reg-

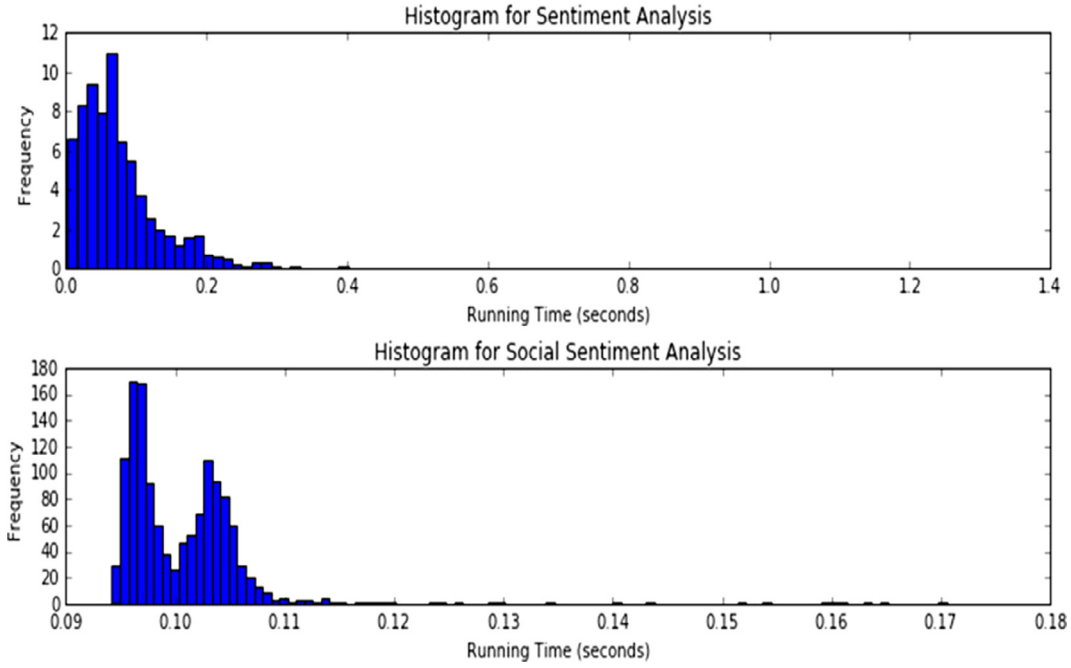

Fig. 15. Running time of nlp vs running time of vader.

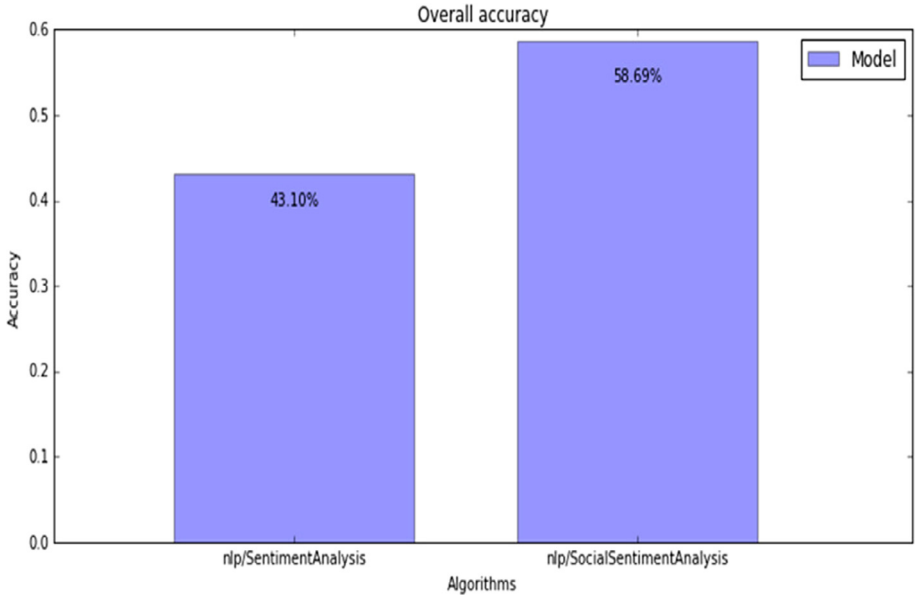

Fig. 16. Accuracy comparison of nlp vs vader.

istered users can. Forums generally only focus on one subject and, therefore, the use of forums as databases ensures that feeling analysis is performed in one domain. The purpose of the reviews is to show that a certain item is effective, so it is a question of a single field. Meanwhile, both companies and potential users benefit from sentimental analysis of reviews. News articles text is usually formal and structured. The use of graphics in news articles is one issue arising from the extraction in this domain. Graphs and figures can sometimes contain information not found in the article's text. Therefore, it will be ignored the use of existing methods.

The use of graphics in news articles is one issue arising from the extraction in this domain. Graphs and figures can sometimes contain information not found in the article's text. Therefore, it will be ignored the use of existing methods. Classification difficulties, language generalization and negation are the major challenges. In addition, recent researchers in a field of sentiment analysis have been attracted to natural language processing tools and require improvement, while some OM or SA algorithms provide good results.

CRediT Author Statement

Krishna Raj P.M: Conceptulization. **Jagadeesh Sai D:** Conceptulization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] V. Jatav, R. Teja, S. Bharadwaj, V. Srinivasan, Improving part-of-speech tagging for NLP pipelines, arXiv preprint arXiv:1708.00241, 2017.
[2] B. Jurish, K.-M. Würzner, Word and sentence tokenization with hidden markov models, J. Lang. Technol. Comput. Linguistics 28 (2) (2013) 61–83.
[3] P. Anupriya, S. Karpagavalli, LDA based topic modeling of journal abstracts, 2015 International Conference on Advanced Computing and Communication Systems, 2015.
[4] L. Huang, J. Ma, C. Chen, Topic detection from microblogs using T-LDA and perplexity, 2017 24th Asia-Pacific Software Engineering Conference Workshops (APSECW), 2017.
[5] P. DiMaggio Adapting computational text analysis to social science (and vice versa) Big Data & Society 2 2 2015 205395171560290 10.1177/2053951715602908
[6] https://www.sacred-texts.com/hin/maha/index.htm
[7] Neeraj Kumar, "Conventional Neural Network based biometric Detection Algorithm", Journal of Control, Measurement, Computing and Communications, (SCI).
[8] D.M. Rayudu "Naresh. E and Dr. Vijaya Kumar B. P, The Impact of Test-Driven Development on Software Defects and Cost: A Comparative Case Study" International Journal of Computer Engineering and Technology (IJCET) 5, no. 2 2014
[9] E. Naresh, B.P. Kumar, Vijaya Kumar, M. Niranjanamurthy, B. Nigam, Challenges and issues in test process management, J. Comput. Theor. Nanosci. 16 (9) (2019) 3744–3747.
[10] Muskan Jindal Eshan Bajal Alakananda Chakraborty Prabhishek Singh Manoj Diwakar Neeraj Kumar 37 2021 2952 2958

Further Reading

[1] J. Mandhyani, L. Khatri, V. Ludhrani, R. Nagdev, S. Sahu, Image sentiment analysis, Int. J. Eng. Sci. (2017).
[2] R. Feldman. Techniques and applications for sentiment analysis" Communications of the ACM 56 4 2013 82 89
[3] V. Gajarla, A. Gupta, Emotion detection and sentiment analysis of images, Institute of Technology, Georgia, 2015.
[4] W.Y. Kim J.S. Ryu K.I. Kim U.M. Kim A Method for Opinion Mining of Product Reviews Using Association Rules 2009 Information Technology, Culture and Human hlm. 270–274.
[5] A. Go, R. Bhayani and L. Huang "Twitter sentiment classification using distant supervision". CS224N Project Report, Stanford, 1: 122009
[6] Qiang Ye, Ziqiong Zhang, Rob Law, Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, Expert Syst. Appl. 36 (3) (2009) 6527–6535.
[7] N. Mittal, D. Sharma, M.L. Joshi, Image sentiment analysis using deep learning, in: In: International conference on web intelligence (WI). IEEE, 2018, pp. 684–687.
[8] Y. Wang, B. Li, Sentiment analysis for social media images, in: In: International conference on data mining workshop (ICDMW). IEEE, 2015, pp. 1584–1591.
[9] E. Baralis, S. Chiusano and P. Garza. (2008). A Lazy Approach to Associative Classification. IEEE Trans. Knowledge Data Engineering. 20(2): 156-171.
[10] Erik Boiy, Marie-Francine Moens, A machine learning approach to sentiment analysis in multilingual Web texts, Inf. Retrieval 12 (5) (2009) 526–558.
[11] A. Agarwal, P. Bhattacharyya, Sentiment analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified, in *Proceedings of the International Conference on Natural Language Processing (ICON)*, 2005.