International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

APPLICATION OF SENTIMENT ANALYSIS IN WEB DATA ANALYTICS

Aleena Rose K.S

Department of Computer Applications, Sree Narayana Guru Institute of Science and Technology, Ernakulam, Kerala, India

---

Abstract - *Sentiment analysis is a predominantly classification algorithm aimed at finding an opinionated point of view and its disposition and highlighting the information of particular interest in the process. The applications of sentiment analysis are broad and powerful sentiment analysis system for text analysis combines natural language processing (NLP) and machine learning techniques to assign weighted sentiment scores to the entities, topics, themes and categories within a sentence or phrase. The Netflix and Stanford models are the two wide cited sentiment analysis models used in this study. The Objective of the Project is to motivate the use of sentiment analysis as a technique for analyzing the presence of various aspects like human trafficking in escort ads pulled from the open web. Old techniques have not focused on sentiment as a textual cue and instead have focused on other visual cues (e.g., presence of tattoos in associated images), or matter cues (specific types of ad-writing; keywords, etc.) Here, we also aimed to implement an automated knowledge discovery on small businesses by using local business websites and Twitter as data source and analyzing the data with data mining techniques. By using this app we can easily choose the appropriate website which contains relevant information. This is done by machine learning techniques and rates those keywords as positive, negative or neutral based on sentiment analysis. Keywords: Data mining, Human Trafficking, Machine Learning, Sentiment Analysis, Web crawling.*

Key Words: **Data mining, Human Trafficking, Machine Learning, Sentiment Analysis, Web crawling.**

1. INTRODUCTION

Human trafficking involves the use of force, fraud, or coercion to obtain some type of labor or commercial sex act. Every year, millions of men, women, and children are trafficked worldwide – including right here in the United States. It can happen in any community and victims can be any age, race, gender, or nationality. Traffickers might use violence, manipulation, or false promises of well-paying jobs or romantic relationships to lure victims into trafficking situations. Enabling intelligent search systems that can navigate and facet on entities, classes and relationships, rather than plain text, to answer questions in complex domains is a longstanding aspect of the Semantic Web vision. The search engine has been rigorously prototyped as part of the DARPA MEMEX program and has been integrated into the latest version of the Domain-specific Insight Graph (DIG) architecture, currently used by hundreds of US law enforcement agencies for investigating human trafficking. Our work helped manufacture indices and internet corpora together with eighty million web content and forty million pictures, and during this vast of information. This team also worked with enforcement clients to collect information and to produce proof and knowledge that saved the lives of many victims of human trafficking. We were mainly interested in mining the data associated with the advertisements, and also in the area of sentiment analysis [8].Sentiment analysis of web data is an approach to discern the text writer's affinity or negativity as expressed through her use of language and vocabulary. Sentiment can be binary or categorical/multiclass and can serve as a data summarization for large bodies of text, social media data, etc. We are not aware of extensive studies of sentiment analysis as it application of human trafficking is wide, and our hypothesis was that sentiment analysis could be an important textual cue which indicates a web document's ability to describe a real trafficking scenario. Sentiment analysis could also help to provide an opportunity to open the mindset of both the person writing the ad-the potential predator; or even the victim. In this research, we try to describe a series of experiments to use sentiment associate degree analysis as a sign for human trafficking. We applied existing binary e.g., Netflix [9] and categorical eg: Stanford Treebank [15] sentiment models on to branches of internet ads from our MEMEX human trafficking corpus that ground-truth concerning human trafficking was on the world. Furthermore we trained 2 more ensemble sentiment analysis models that used 2 categorical models, and incorporated extra options as well as geographic location as known through named entity extraction [11] and also the presence of negation within the text as additional cues. The ensemble models performed well than individual models in accuracy and variety of iterations needed to bear down in. This paper also focused to work out an automatic data discovery on little businesses by victimization native business websites and Twitter as data sources and analyzing the info with data processing techniques. We tend to outlined websites that contain info regarding local little businesses as stable information supplies and Twitter as a radical information resource. We first tried to get keywords regarding little businesses and use the keywords for search queries on Twitter. To research clients' comments obtained from the questions, we tend to devise text mining techniques to work out what quantity positive or negative the comments mean. To utilize user info, WHO wrote the obtained comments; we tend to outline a list of keywords that may be clues for users' identities like age, gender, and

occupation. Lastly, this information were used to create linguistics info about user preferences on little businesses and aggregate to supply helpful info for important requests from little business house owners or dynamic QA systems which provide edifice recommendation services.

2. RESEARCH AREA

Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs. Data mining depends on effective data collection, warehousing, and computer processing. The data mining process breaks down into five steps. First, organizations collect data and load it into their data warehouses. Next, they store and manage the data, either on in-house servers or the cloud. Business analysts, management teams and information technology professionals access the data and determine how they want to organize it. Then, application software sorts the data based on the user's results, and finally, the end-user presents the data in an easy-to-share format, such as a graph or table. Data mining programs analyze relationships and patterns in data based on what user's request. For example, a company can use data mining software to create classes of information. To illustrate, imagine a restaurant wants to use data mining to determine when it should offer certain specials. It looks at the information it has collected and creates classes based on when customers visit and what they order. In other cases, data miners find clusters of information based on logical relationships or look at associations and sequential patterns to draw conclusions about trends in consumer behavior. Warehousing is an important aspect of data mining. Warehousing is when companies centralize their data into one database or program. With a data warehouse an organization may spin off segments of the data for specific users to analyze and use. However, in other cases, analysts may start with the data they want and create a data warehouse based on those specs. Regardless of how businesses and other entities organize their data, they use it to support management's decision-making processes. A web crawler (also known as a web spider or web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner. This process is called Web crawling or spidering. Many legitimate sites, in particular search engines, use spidering as a means of providing up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for spam).Web search engines and some other sites use Web crawling or spidering software to update their web content or indices of others sites' web content. Web crawlers copy pages for processing by a search engine which indexes the downloaded pages so users can search more efficiently. Crawlers consume resources on visited systems and often visit sites without approval. Issues of schedule, load, and" politeness" come into play when large collections of pages are accessed. Mechanisms exist for public sites not wishing to be crawled to make this known to the crawling agent. For example, including a robots.txt file can request bots to index only parts of a website, or nothing at all. The number of Internet pages is extremely large; even the largest crawlers fall short of making a complete index. For this reason, search engines struggled to give relevant search results in the early years of the World Wide Web, before 2000. Today, relevant results are given almost instantly. Machine learning (ML) is the study of computer algorithms that improve automatically through experience. [1] It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. [2] Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. [4][5] In its application across business problems, machine learning is also referred to as predictive analytics. Artificial intelligence is a technology that is already impacting how users interact with, and are affected by the Internet. In the near future, its impact is likely to only continue to grow. AI has the potential to vastly change the way that humans interact, not only with the digital world, but also with each other, through their work and through other socioeconomic institutions – for better or for worse. If we are to ensure that the impact of artificial intelligence will be positive, it will be essential that all stakeholders participate in the debates surrounding AI.

3. RELATED WORK

Most of the state of the art works and researches on the automatic sentiment analysis and opinion mining of texts collected from social networks and micro blogging websites are oriented towards the classification of texts into positive and negative. In this paper we present our machine learning experiments with regard to sentiment analysis in blog, review and forum texts found on the World Wide Web and written in English and other languages. We train from a set of example sentences or statements that are manually annotated as positive, negative or neutral with regard to a certain entity. We are interested in the

feelings that people express with regard to certain consumption products. We learn and evaluate several classification models that can be configured in a cascaded pipeline. We have to deal with several problems, being the noisy character of the input texts, the attribution of the sentiment to a particular entity and the small size of the training set. We may succeed to identify positive, negative and neutral feelings to the entity under consideration with ca. 83 percentage accuracy for English texts based on unigram features augmented with linguistic features. The accuracy results of processing the Dutch and French texts are ca. 70 and 68 percentage respectively due to the larger variety of the linguistic expressions that more often diverge from standard language, thus demanding more training patterns. In addition, our experiments give us insights into the portability of the learned models across domains and languages. A substantial part of the article investigates the role of active learning techniques for reducing the number of examples to be manually annotated. Boiy and Moens perform sentiment analysis on open web data using Apache OpenNLP. They focus on training three-class (positive, negative, and neutral) sentiment related to consumer products on text parsed from blog, review and forum sites. The authors were able to leverage their approach on English, Dutch and French text, with 83 percentage accuracy on the English texts. In their text analysis, the authors leverage unigram features-similar to our own approach. Unlike our approach, however, the authors do not mention how they isolated the appropriate text from the blog, review and forum sites, and performed HTML pre-processing, or if they used Text to-Tag (TTR) techniques as we employed. Others have also used Apache OpenNLP for sentiment analysis including the Elixa project, the work by Wogensteinetal and Johnsonetal. As noted by Paltoglou, no models are widely available however for sentiment analysis using Apache OpenNLP. Tim Weninger and William H Hsu present Content Extraction via Tag Ratios (CETR) – a method to extract content text from diverse webpages by using the HTML document's tag ratios. We describe how to compute tag ratios on a line-by-line basis and then cluster the resulting histogram into content and non-content areas. Initially, we will find that the tag ratio histogram is not easily clustered because of its one dimensionality; therefore we extend the original approach in order to model the data in two dimensions. Next, we present a tailored clustering technique which operates on the two-dimensional model, and then evaluate our approach against a large set of alternative methods using standard accuracy, precision and recall metrics on a large and varied Web corpus. Finally, we show that, in most cases, CETR achieves better content extraction performance than existing methods, especially across varying web domains, languages and styles. Richard Socher, Alex Perelygin and their team proposes a research article on Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. They introduced Recursive Neural Tensor Networks and the Stanford Sentiment Treebank. The combination of new model and data results in a system for single sentence sentiment detection that pushes state of the art by 5.4 percentages for positive/negative sentence classification. Apart from this standard setting, the dataset also poses important new challenges and allows for new evaluation metrics. For instance, the RNTN obtains 80.7 percent accuracy on fine-grained sentiment prediction across all phrases and captures negation of different sentiments and scope more accurately than previous models.

4. PROPOSED WORK

In searching a pathway to apply sentiment analysis to our agency MEMEX human trafficking connected web content, we decided to apply 2 prevailing model classes :(1) binary sentiment-positive/ negative; and (2) categorical, or multiclass sentiment-e.g. ,like, love, neutral, etc. We tend to start our approach by analyzing wide cited and used binary and categorical models for sentiment analysis that square measure either pre-trained on various existing net and social media knowledge, or that offer the first data and allow the user to show their own coaching. We tend to dim our search right down to the Netflix binary sentiment and Stanford Treebank categorical sentiment models.

4.1 The Netflix

The 33.1MB Netflix dataset include the information which was collected through the supply cited. This supply delivers the positive and negative records in 2 different directories, which allowed U.S. to offer the whole text record in an individual label (positive or negative). We decided to merge the 2 directories together in to one coaching dataset. Though this data set isn't human trafficking only, it's wide employed in the machine learning community and a basic model for sentiment supported web-based matter cues and ads.

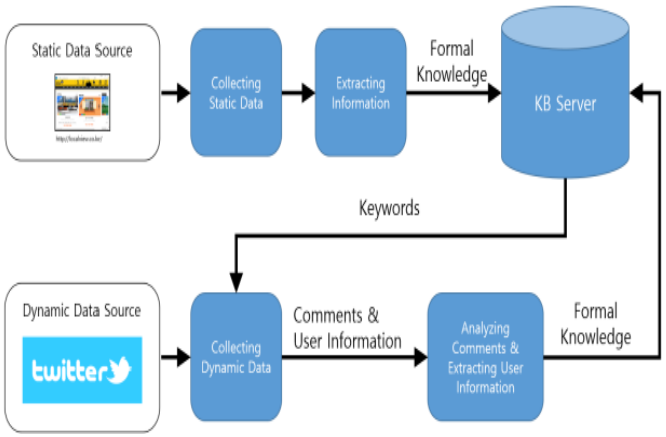
4.2 The Stanford

The 345MB Stanford Tree bank knowledge includes a dataset which was collected through the supply cited. The each JSON based format was given a numerical score from zero to one linking matter options to a score. We have an influence to utilize the score and divided the data into variety of classes comparable to the recent Facebook's reaction to free sentiment options. Facebook reactions square measure multi-class sentiment labels to connect to text. We then planned to use Stanford's Treebank during a similar categorical sentiment analysis and also took all their views with a score.

4.3 System Architecture

The whole processes for the data detection area unit reviewed within the overview of knowledge discovery process. It includes internet crawl processes from static and dynamic knowledge resources on the net, data extraction and analysis processes from every knowledge supply, and so the cognitive content server, that contains the data information.

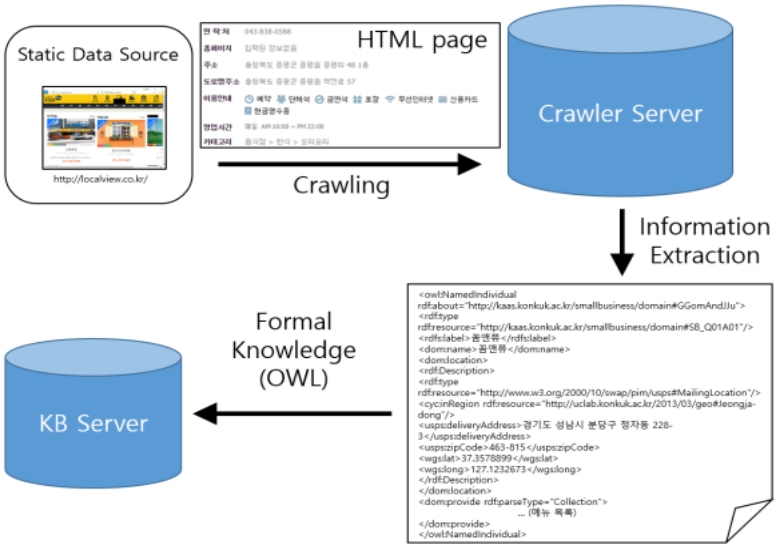
Fig -1: Overview of the knowledge discovery process



4.4 Static Data Collection

Among different types of data on small businesses, we framed data that are not formed by users and likely to keep unchanged for a long period as static data. These data include their names, locations, contacts, types of businesses, etc. To produce static data of small businesses, we used localview.com as the resource. We plan to conduct web crawling on the website, produce formal knowledge in OWL by preprocessing the obtained data, and added the OWL to the knowledge base server.

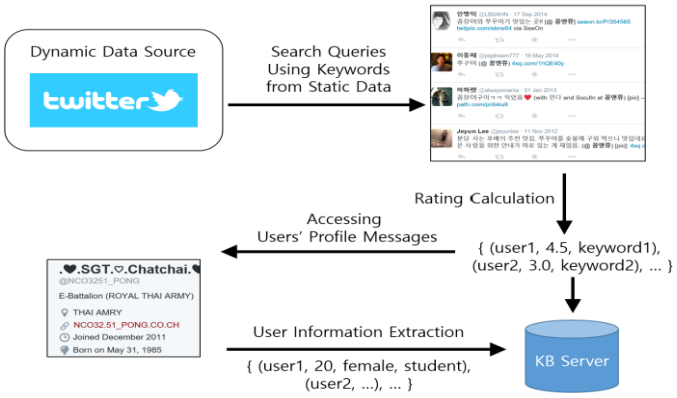
Fig – 2: Static Data crawling process



4.5 Dynamic Data collection

We also derived another type of data called dynamic data, which are obtained by users and most probably to change over time than static ones. This includes businesses' reputations for their services. We plan to choose Twitter as the dynamic data source. We planned to use businesses' names obtained from the static data source as keywords to do search queries on twitter, and collected tweets which includes the keywords and profile messages of users who wrote the tweets. To produce fruitful results, there is a need to perform extra processes including text mining techniques into the process that have already proven successful.

Fig – 3: Dynamic data crawling process



4.6 Research Analysis

Each message in the twitter containing businesses' names can reflect their opinion about the business. However, as for Tweeter, clients' comments just come in words without any ratings. So some kind of analysis is needed to find out how much positive or negative each of the tweets is about their businesses. To solve this challenge, we plan to innovate a method to measure how much positive a user's replies on social platforms. The first step is to collect users' comments on websites with comments' ratings followed by splitting each comment by word and count the number of each word occurring in the comments. At the same time, for each word, it is necessary to add up the ratings of comments where the word appeared. After identifying the number of occurrences and the sum of the ratings of each word, dividing the latter by the former produces the average rating for each word. The words that often appear in positive or high-rating reviews show high average ratings, and words that often appear in negative or low-rating reviews show low average ratings as usual. With the help of table consisting of words and their average ratings, it is possible to measure ratings of any comments. For instances, if the maximum rating is 5.0, then the median value is to be 2.5. As a result, the more positive words a comment has, the more rating it gets.

Table -1

Word	Rating
love	4.0
tasty	5.0
but	2.0
pricey	1.0
...	...

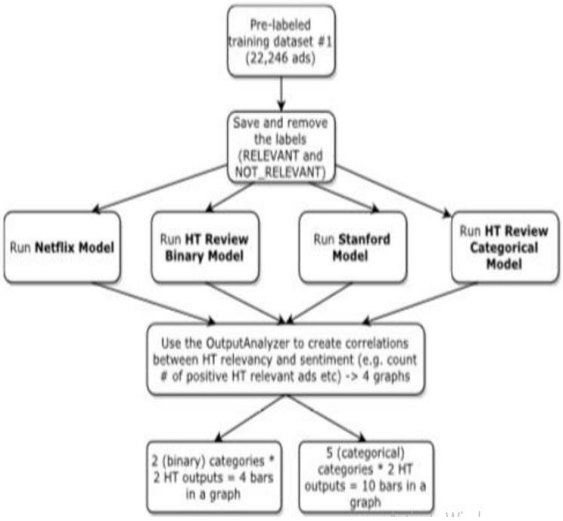
4.7 Extraction of Data

Besides users' opinions, information about the users like age, gender, occupation are also a great resource to support small businesses' decision making in their marketing or operating strategies. Users on Tweeter usually show these pieces of information on their profile messages. But these are just written in natural language sometimes with their own self-expressions. By observing a lot of the profile messages of Tweeter account holders, we could identified several keywords which could figure out their personal data including gender, age, and occupations from their profile messages by exact keyword matching followed by exclusion of their profile messages which do not contain any of the keywords. Each user's information forms a pair with his or her tweet's rating calculated by the previous step, and are stored in the knowledge base server.

4.8 Approach

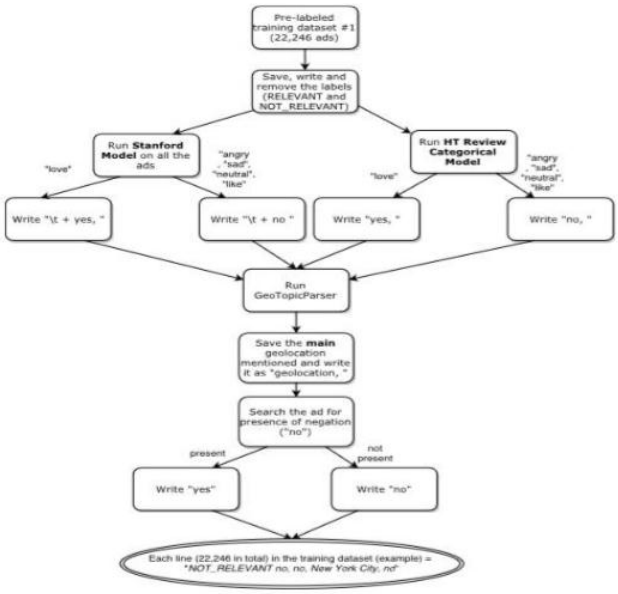
Our overall perspective of this study is divided along our training models and testing models. The first part of our approach were to explore correlations between off- the-shelf sentiment models Netflix, Stanford Binary/Categorical, and HT Provider Review Binary/Categorical and their outputs and the ground truth labels from the HT ground truth dataset. The steps of this portion of the approach are evaluating trained off-the-shelf sentiment models and HT Provider review models

Flow Chart-1



First and foremost part of our analysis uses the HT ground truth pre labeled (RELEVANT, NOT RELEVANT) dataset, runs the training models and measure the number of positive and negative ads per each of the two initially given labels. The goal is to analyze the distribution of sentiment among the HT-relevant and not relevant ads and create correlations. This analysis explains some basic details on the presentation of the models and how they differ depending on many factors. The HT categorical model is the backbone of our hypothesis about the importance of love label in analyzing HT data. Based on its classification, the love label shows a good sign of human trafficking relevancy, and angry is a good indicator of human trafficking irrelevancy. The HT Ground Truth Binary model has a general tendency to classify more ads as NOT RELEVANT compared with the HT Ensemble I which produces many more RELEVANT classifications. After performing exploratory analysis on resulting identified correlations between our training models and test HT Ground Truth data, we plan to determine that the HT Categorical model and Stanford categorical model will provide correlations with HT RELEVANT ads when the sentiment analysis from both models labeled the ads as love. Furthermore, as previously noted, we also may produce a high degree of RELEVANT HT ads exhibited repeating geo locations (example Las Vegas). In the final step, the ad text in HT RELEVANT ads plan to use extreme negative words and language ("negation"). Given this, we will follow with ensemble sentiment model generation approach.

Flow Chart- 2



5. CONCLUSIONS

The project was successfully completed within the time span allotted. Every effort has been made to present the system in more user friendly manner. All the activities provide a feeling like an easy walk over to the user who is interfacing with the system. A trial run of the system has been made and is giving good results. Based on early work in this area sentiment analysis is a viable classification mechanism and proxy to identify human trafficking in web data. Using open source sentiment models and models trained on human trafficking provider review data we were able to use exploratory analysis to find trends suggesting an approach for what textual, sentiment, geographic and natural language cues are appropriate features to indicate if an ad is trafficking or not and to build accurate ensemble models to automatically identify it. It can be concluded that the HT Ensemble I Model, with a training set accuracy of 0.84 at iteration 100 and test set accuracy of 0.52 at iteration 100, clearly outperforms its competitors. We also studied about building a small business knowledge base in an automated way by exploiting data on the web. We first defined two types of data: static data and dynamic data. For the static data, we collected data from a local website containing information about small businesses. And for the dynamic data, we used Twitter and collect tweets and users' information through open API it provides. What we aimed to collect from the dynamic data is businesses' reputations among online users. However, users' comments on social media usually do not have ratings to indicate how much positive or negative their reactions are. So we devised a method to measure ratings of comments. We used comments on menupan.com with ratings on restaurants. By using words' ratings obtained from the method, we could calculate ratings of comments from Tweeter. We also showed correlation coefficients of the method's results against the users' actual ratings on the website.

REFERENCES

1. https://link.springer.com/article/10.1007/s10791-008-9070-z
2. https://www.investopedia.com/terms/d/datamining.asp
3. M. Anastasija, A. Chris. Mattmann, "Ensemble Sentiment Analysis to Identify Human Trafficking in Web Data," (2017).
4. B. Mondher and O. Tomoaki, "Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter," In Communications (ICC), IEEE International Conference on. IEEE, 1–6.2015.
5. J. Christopher, S. Parul, and S. Shilpa, "On classifying the political sentiment of tweet," Cs. utexas. Edu (2012).
6. K. Kamala, S. Jyoti, and P. Bandana. $ Trafficking and prostitution reconsidered: New perspectives on migration, sex work, and human rights. Routledge, 2015.
7. K. Mikhail, S. Nikunj, and Kiran Vodrahalli, "A Large Self Annotated Corpus for Sarcasm," ArXiv preprint arXiv: 1704.05579 (2017).
8. Nicholas D Kristof, "Where pimps peddle their goods," The New York Times 11 (2012).
9. L. Bing Liuand, "A survey of opinion mining and sentiment analysis," In mining text data. Springer 415–463. 2012.
10. Andrew L Maas, Raymond E Daly, Peter T Pham, DanHuang, Andrew Y Ng, and ChristopherPotts, " Learning word vectors for- sentiment analysis," 2011.
11. C. Mattmann. [n.d.]. GeoTopicParser. http://wiki.apache.org/tika/GeoTopicParser.
12. A. Chris Mattmann and Madhav Sharan, "An Automatic Approach for Discovering and Geocoding Locations in Domain-Specific Web Data," In Proceedings of the 2016 IEEE 17th International Conference on Information Reuse and Integration, 87–93. 2016.
13. A.Mensikova and C.Mattmann.[n.d.]. USC Data Science-Human Trafficking Lead Generation Analysis. http://irds.usc.edu/Sentiment AnalysisParser/htlg. html.([n.d.]).
14. ApacheOpenNLP. Apache software foundation URL: http://opennlp.apache. org (2011).
15. InakiSanVicente, S. Xabier, and A. Rodrigo, "Elixa: Amodular and flexible absa platform''. ArXiv preprint arXiv: 1702. 01944 (2017).
16. Tim Weninger and William H Hsu, "Text extraction from the web via text-to-tag ratio In Data base and Expert Systems Application'', DEXA'08. 19th International Workshop on.IEEE, 23–28. 2008
17. W. Florian, D. Johannes, Dirk Reinel, Sven Rill, and Jorg Scheldt, "Evaluation of an algorithm for aspect-based opinion mining using a lexicon based approach." In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining. ACM, 5.2013

AUTHOR

Aleena Rose K.S currently pursuing the Dual degree MCA with MG University, Kottayam, Kerala.