Modeling a Functional Engine for the Opinion Mining as a Service using Compounded Score Computation and Machine Learning

Rajeshwari D[1]
Research Scholar, ATMECE, Assistant Professor,
Department of Information Science & Engineering,
NIE Institute of Technology, Mysuru, India

Puttegowda.D[2]
Professor & Head, Department of Computer Science &
Engineering, ATMECE, Mysuru, India

***Abstract*—The ever-growing use of the digital platform for the various walks of the applications, primarily on the collaborative platforms of e-commerce, e-learning, social media, blogging, and many more, produces a large corpus of unstructured text data. Many potential strategic solutions require an accurate and fast classification process of the Opinion's text corpus hidden patterns. In-premise applications have various real-time feasibility constraints. Therefore, offering an Opinion as a Service on the cloud platforms is a new research domain. This paper proposes a design framework of the evolution of the classification engine for opinion mining using score-based computation using a customized Vader algorithm. Another method for scalability is a machine learning model that supports a large corpus of unstructured text data classifications. The model validation is performed for the various complexes, unstructured text datasets with the different performance metrics of the cumulative score, learning rate, loss function, and specificity analysis. These metrics indicate the models' stability and scalability behaviors and their accuracy and robustness across different datasets.**

***Keywords*—Text mining; opinion; sentiments; machine learning; unstructured data; cloud services**

I. Introduction

The evolution of web2.0 and Cloud has brought a complete change in the digital system's development and production [1]. Global resource constraints and economic liberalization lead to realizing a collaborative business model. A highly distributed production-distribution and consumption market require an ecosystem of technology that has high availability and scalability—Cloud computing service offerings cater to these demands [2]. The competitive environment of cloud service providers (CSP) and the enterprise demands various services apart from the Cloud's traditional offerings. The evolution of the words' representation into vectors provides an ease to process the word corpus and leads a technology, namely text analytics. Various open platforms offer a facility to express the feedback or textual expression in many contexts of the brand-building process, marketing, or product campaign. The corpus of the text contains the hidden treasure of the Opinion. It is not economically feasible for the individual organization to set up dynamically evolving methods for the opinions mining as in-premise computing infrastructure. Therefore, the CSPs are in the process of building an ecosystem to offer Opinion-Mining as a Services (OMaaS). This paper proposes an architectural model for the Opinion-Mining design as a Service (OMaaS) offering from the CSPs. The basic workflow diagram of the 'OMaaS' is as in Fig. 1.

The framework for the OMaaS provisions a system to acquire the Cloud users (CS) text corpus (Tc) through a dedicated channel with the dashboard of the virtual layer (VL) to the cloud data store. It handles the large corpus that further gets synchronized to the cloud data text analytics Engine (TAE), where the opinion mining's effective algorithm gets executed. Finally, the respective CSi ∈{CS} gets the visual or statistical representation of the mined Opinion from the respective Tc. Such a model's overall success largely depends upon how effective, and in a scalable manner, the view is mined on a real-time basis.

Many ubiquitous applications are conceptualized, where text analytics plays very crucial roles. Many of such application may include: i) Dynamic info-system on the dashboard of the vehicles, ii) business strategic decision tools, iii) topic modeling, iv) summarization, v) patent data matching, vi) health care decision support system, vii) the forensic tool, viii) decision making based on feedback – sentiment analysis, ix) political campaign, x) historical literature analysis, xi) visual search. Section II describes various researches that took place in the field of text analytics in a different context. Section III provides the descriptions of the diverse dataset taken into consideration for the model variation followed by the Sections IV and V for the two respective models of cumulative score and machine learning-based classification algorithms as a proposed engine Opinion mining to be synchronous with the OMaaS. Finally, Section VI discusses the results and analysis, followed by a conclusion in Section VII.

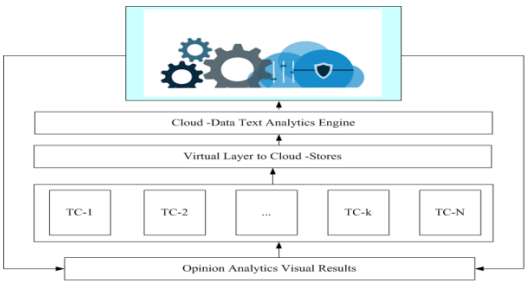

Fig. 1. Workflow Process Diagram of OMaaS.

II. Review of Literature

The use of text-data from social media like Facebook and other feeds and surveillance camera images (**Neuhold et al., 2018**) is found. The text analytic is exploited to display the road condition on the dashboard [3]. The big-data and the business complement's process management complement if the text generated is adequately analyzed [4]. A tree-based visual representation of text is being practiced, but this method is not scalable [5]. In research, the usual challenges to the researchers are to handle a large query result. The bag of words, along with natural language processing and visual analytics, is being studied in the work of **Benito et al., (2019)** [6]. **Basole et al., (2019)**, has reviewed topic modeling on an extensive text description used in a business domain based on text analytics [7]. Summarization or building abstraction of a large document is beneficial to grab quick knowledge. Text analytics is used by **Han et al., (2020)** in their work [8]. **Reghupathi et al., (2018)**, has examined the use of text analytics on patent data corpora using a concept like a word count and co-occurrence and the machine learning model [9]. Health care industries are another domain which produces a vast amount of unstructured text data (**Kumar et al., 2019**), has performed text analytics for decision support system [10].

The use of sequence-to-sequence learning in text analytics is becoming popular to build many text analytics-based applications with higher accuracy and lower training time (**Keneshloo et al., 2019**) [11]. Many giants like Facebook and Google use text analytics for their respective goals. Similar benefits can be achieved in further education, banking, and marketing sectors [12]. The forensic sector benefits if the complex text data from various communication sources are being analyzed (**Koven et al., 2019**), devices a tool that uses text analytics on the email data corpus [13]. Nowadays, the topic modeling algorithm is gaining popularity (**EI-Assady et al., 2018**), proposes a decision-making technique based on relevant feedback using text analytics [14]. The study of sentiment analysis in crowdfunding is presented by (**Wang et al., 2017**) [15]. Media is another domain where a large corpus of text data is generated. Text analytics facilitates benefits on the topic description of an event as in the work of (**Lu et al., 2018**) [16]. Text analytics has also shown its benefits in the political election campaign (**Gad et al., 2015**), proposes an analytics tool for the visual representation of the social message trend [17].

The analysis of semantic with its content plays a vital role in content analysis [18]. **Ojo et al., (2019)**, present patient sentiment analysis using textual data [19]. **Karam et al., (2016)**, proposes a design of new hardware that supports the ecosystem of processor and memory for test analytics [20]. **Vatrapu et al., (2016)**, explores set theory-based visualization to complement text analysis [21]. The sedimentation-based visualization concept of coordinated structure in text analysis has been studied by **Liu et al., 2016)** [22] and **Sun et al., (2016)** [23], respectively. Different regional history analysis is possible by text data analysis such study for Roman history is being carried out in the work of **Cho et al., (2016)** [24], various web-based visualization tool and fundamentals of visual text analytics are described in the work of **Liu et al., (2019)** by analyzing a large corpus of published papers using concurrence relationship [25]. The basic features like parts of speech, text color, and font size make the corpus complex; an extensive survey is being conducted by Strobelt et al., (2106), different understanding highlighting, and visual search techniques [26]. In most text analytic methods, structuring the respective word with their meaning is crucial to arrive at an efficient qualitative and quantitative representation to achieve accuracy like a human [27].

III. Dataset Description

The OMaaS framework proposes two core models for the classifications, which use the following datasets for evaluating the algorithms for the text analytics engine for the opinion classifications: i) Partial Complex Text and emojis, ii) fastText Facebook's AI Research (FAIR) lab[28], iii) Opinion Data from the University of Illinois, Chicago[29]

IV. Modelling a Complex Content: Hybrid Opinion using Text and Symbol using Customized Vader Algorithm

A. Vector of Text Token (TTo)

The simulation environments are controlled by initializing a Mersenne Twister generator with seed '0' [30]. The system deals with the complex heterogeneous constructs using text token and the symbols as **Cf** ={**T**∪**S**}, where T= text token and S= symbols, as nowadays it is a fashion that people express their statements or Opinion with the combined format of the text sentence partially and complement it with some symbols (shown in Table I).

TABLE I. Illustrates Some Typical Examples of the Construct of Such a Dataset

SL.	Construct	Real meaning
1	in office 😟 wait #weekend 😇	Bored in the office waiting for the weekend
2	#weekend 😇😇😇	Becoming happy for the weekend

The algorithm 1 is described below:

Algorithm 1: Generating Vector of Text token from Complex format(unstructured)

Input: Cf
Output: TTo
Process:
Start
Initialize Cf ←Fn
tCf←*f1*(Cf : ∀ content ∈Cf)
Tokenization:
TTo ←*f2*(∀ (Td) ∈*tCf)*
End

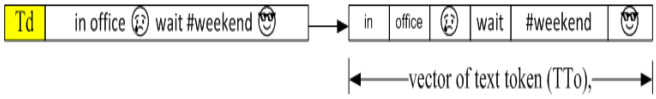

Fig. 2. TTo: Vector of Text Token ∀ Text Data (Td) ∈tCf.

In a start-up, the file containing the complex inputs of 'T' and 'S' as **Fn** gets assigned in Cf's initialization. The 'Cf' transformation takes place into the tabular format for ease of computation, making the characteristics of ∀ content ∈Cf type: 'character' (**Ch**), and tCf represent the transformed format**.** To generate a vector of text token (TTo), the ∀ text data (Td) ∈*tCf* passes through a function of the document tokenization process (f2()), as shown in Fig. 2.

B. Score Weightage

The score weightage (**Sw**) for ∀ tokenized-Document (**tD**) ∈TTo is computed using a popular "*valence aware dictionary*" for sentiment reasoning: "Vader" as customized Vader(**cV)** [31]. The large corpus of unstructured textual data transformation and gaining a quantitative ratio are the engine's main goals. In future applications, artificial intelligence (AI) based service are depicted as 'OaaS' models in the enterprises' CRM applications' intrinsic parts. The cV is basically a rule-oriented lexicon model (**ROLM**) based on the set of {**sL**, **gR**,**sYC**}, where, sL= 'sentiment lexicon, gR= 'grammatical rules and **sYC**= 'convention', such that sYC:→{**s.P**,**s.I**}, such that s.P= polarity , s.I=intensity. The cV constructs a 'wordlist' with the wide-ranging list of feature-vector (Fv) such that Fv={Word(W), Phrases(P), Emo-icons (Ei), Acronyms (Ac)} with the rating of s. P and s.I in a **-ve** score to the **+ve** score, and the average is assigned as Sw. Vader's customization involves the handlers for the other parts of speech, characterization, and punctuations. The cV takes the entire *tCf*and their associated Fv and operates on {**s. P**,**s. I**} as per the specific rule sets. Finally, the summation of all the Fv scores gets normalized by scaling it in the range: **R** [-1 to 1] using equation (1) for the compounded score, Sc.

$$Sc = \frac{S}{\sqrt{S^2+\beta}} \ldots \quad (1)$$

Algorithm 2: Custom Vader algorithm for computing compound score Sc

Input: TTo

Output:*Sc*
Process:
Start
cV ←{sL, gR,sYC}
sYC:→{**s**.P,s.I}
Fv={(W), (P), (Ei), (Ac)}
cV← {*tCf* , Fv } :→ {**s.**P,s.I} as per the specific rule sets
Normalize, Fv by scaling:**R** [-1 to 1]
➔ using $Sc = \frac{S}{\sqrt{S^2+\beta}}$
Update: Sc
End

The value of β approximates the maximum probability of the expected cost of the score S. The algorithm is explained in algorithm 2, and the algorithm is implemented into two distinguished data set to measure the compound scores and the time computations. The results are described in Section VI of the results and discussion.

V. Modelling Complex Content: Machine Learning

A. Auto Label Annotation for Data Model

The artificial intelligence research group (**FAIR**) by Facebook provides a model for creating a vector depiction of the equivalent word as a library. This library is popularly known as '**fast-Text**' and is used to learn text classification by different machine learning models (**MLM**). The system model takes the dataset provided by the University of Illinois, Chicago, namely: 'Opinion-Lexicon (**OL**),' which contains {6789} word list of both Class: {Negative (**Pw**), Positive (**Nw**)} as a text token (**Tt**) [32] sorted in the sequence of **a→z**. Further, a pre-trained model, namely, {'Word-Embedding'} provides an object named Dictionary (**Dc**) containing 9,99,994 tokens of words as string [33].

Algorithm 3: LabelAnnotation Data for Learning Model

Input: OL
Output: Dla
Start:
[Pw / Nw] ←*f*1(OL)
(W) mx 1←Pw ∪ Nw
La (Undefined: La) ←*f*2 ((NaN) m, 1)
CLa←La (Undefined: La): W
D_{la}←W ∪ CLa
End

The explicit function *f*1() takes OL as an input argument, check the correctness of the files and convert ∀ tokens (Tt) ∈ {Negative (Pw), Positive (Nw)}as a string and the concatenation of Pw ∪ Nw, generates a list '**W**' of size m x n, where n=1. Since the W ∈ {String Datatype}, therefore it is characterized as 'Not a Number (Nan),' a function *f*2() converts a list of 'Nan' of size (m x 1) into a list of Categorical variables to store the label annotation (La) for ∀Tt ∈ W. Further, the corresponding elements of the W are mapped:→ La as a categorical Labels (CLa) annotation. The pair of (W and CLa) provides Labelled Annotated data ($\mathbf{D_{la}}$) used for the Learning models.

B. Token-based Filtering

The token-based filtering takes the Labeled Annotated data table (D_{la}) from the previous procedure of Auto Label Annotation for Data Model. The explicit function f3 () takes the $\mathbf{D_{la}}$as an input argument to return the tokenized documents (D_{toc}), which is a set of {T1, T2, Tk, Tn}, where possible as per the text dataset T1 to Tn could be ∈ {#, , www.address.com,}. The process of the function f3() removes the stop words (SW) and also executes the process of stemming [34] or lemmatization [35]. Further, an additional argument passed to the f3() provides the BoW, which can be extended to multi-lingual analysis. Additionally, all the Unicode punctuations or symbols get eliminated after passing the Dtoc into the function designed to remove it. The English language has approximately 225 stop words eliminated from the updated Dtoc after passing a function that handles these stop words as a noise before further processing the text analytics. Finally, the noise processed Dtoc transforms into lower cases for further processing.

Algorithm 4: Token-based filtering

Input: $\mathbf{D_{la}}$
Output: Dt
Start
$D_{toc} \leftarrow f3(\mathbf{D_{la}})$
Update:
$D_{toc} \leftarrow$ punctuation removal (D_{toc})
$D_{toc} \leftarrow$Eliminate stop words (D_{toc})
$D_{toc} \leftarrow$Capitalized lower (D_{toc})
Update: D_{toc}
end

C. SMO based Support Vector Machine Classifier

The SMO based support vector machine classifier (SVMC) creates a dictionary(D) object from the pretrained fastText[28] word model(fTW). The fTW is a training model-1 which has already taken T1 time, and whenever a new dataset needs to be trained so if the transfer learning model[36] is used, then for a new training model as training model-2 takes T2 time, which is lesser time as T2< T as T=T1+ T2 as in Fig. 3.

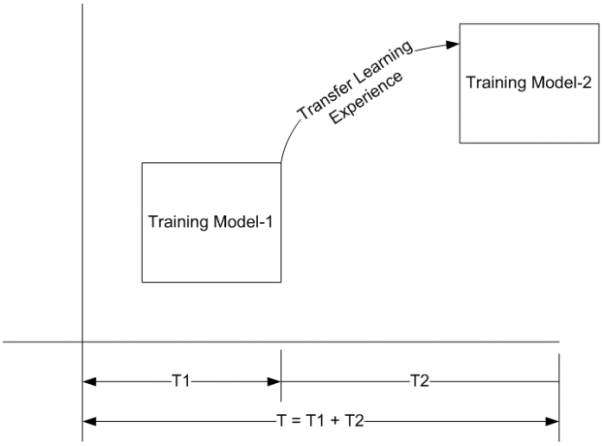

Fig. 3. Time Minimization for Training using a Transfer Learning Approach.

Algorithm 4 describes the process of the SMO-based support vector machine classifier building steps.

Algorithm 5 : Text Classifier Learning Model

Input: D,
Output: tModel
Process:
D←fTW
Call Algorithm-3
Dla(W, CLa)←OL
Check:
Idx← not [dla(W) ∈ D]
Ds←num of W
Random Partition: Cross Validation
[D-train,D-test]←*f*(Ds.numW)
D-train←Word2vector[D-train]
Td←[D-train **U** Cla]
Train SVM- 1-Class-Binary Classifier
tModel←F(Td)

The algorithm 3 LabelAnnotation Data for Learning Model provides Dla ={W, CLa}, further the indexes of all the words Dla(W), which does not belong to the D created as Idx. The total data size **Ds** is the total number of word count **numW**. The Ds partitioning occurs for cross-validation as a random partition on Ds to define the partition for a statistical model ({D-Train, D-Test). The mapping processing of the words to the vector is an essential technique in NLP, which uses ANN to learn a large corpus of the text data, where every word is represented as a list of numbers as a vectorizing simple mathematical function that maps to a semantic similarity as D-train←Word2vector[D-train] and finally the training input to the SVM classifier is obtained as [D-train **U** Cla] → Td. With the Td, the support vector machine (SVM) classifier for one-class and binary classification is trained to get the text classifier model as tModel←F(Td). Fig. 4 and Fig. 5 show the hyper-parameter optimization results status.

The model t is trained on the low-dimension(Low-D) predictors by mapping the independent variables as predictors using a kernel() and support sequential -minimal optimizer(SMO) using an iterative-Single data kernel function or L1- sofmargin minimization whose adjustments with every cycle is shown in Fig. 4 and 5. The confusion matrix for a different dataset for the test performance is described in the results section.

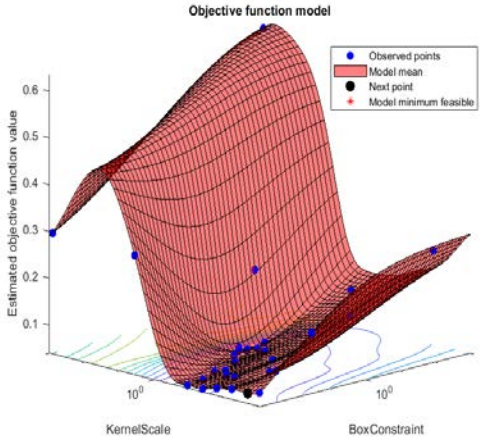

Fig. 4. Objective Function Model.

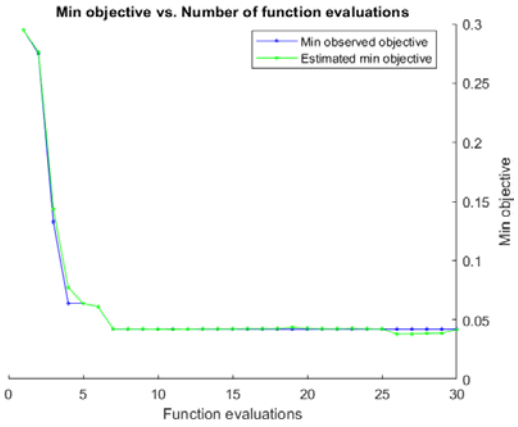

Fig. 5. Minimum Objective vs. Number of Functions Eval.

VI. Results and Discussion

Fig. 6 and Fig. 7 represents the normalized compound scores for TTo with the number of token N= 60 and 1,60,0000, respectively.

The time of processing, including all the process of tokenization, score computation, and visual presentation, is tabulated in Table II below:

It is seen that when the dataset with 50 statements, each average statement time has taken is 3.8 seconds. In contrast, when evaluated on the complex text corpus of 160,0000 views, then the average time taken is 991 sec. Therefore, the consistency y is not maintained as the method is entirely rule-based, and the complexities of the text corpus also vary. For the scalability test, when the same dataset of 50 statements is made multiple copies of 50x, then the time to process is shown in the Table III and its variance as in Fig. 8.

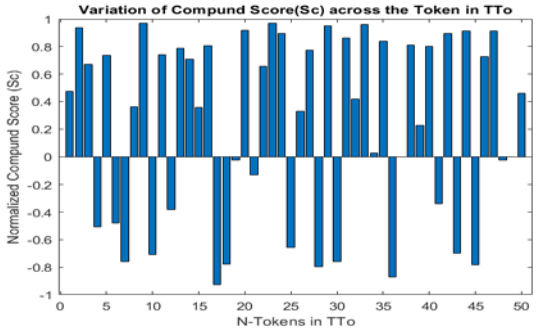

Fig. 6. Variation of Normalized Compound Score Sc across N-Token in TTo, N= 50.

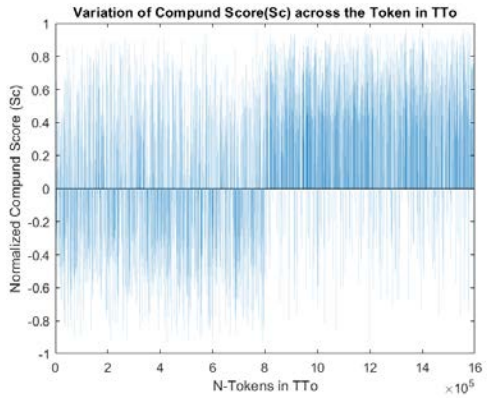

Fig. 7. Variation of Normalized Compound Score Sc across N-token in TTo, N= 160, 0000.

TABLE II. Processing Time for a Small and Large Complex Dataset

Sl. No	Size of the Token	Time to Process (in Sec)
1	50	13
2	1600000	1614

TABLE III. Processing Time for the Scalability Test of a Uniform Dataset

Sl. No	Size of the Token	Processing (in Sec)
1	50	0.33

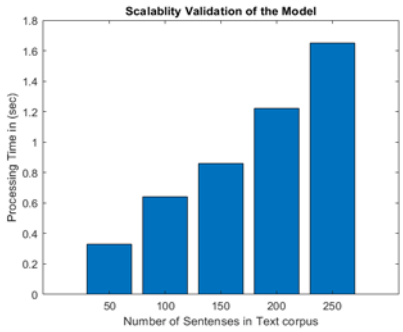

Fig. 8. Variance of Time to Process.

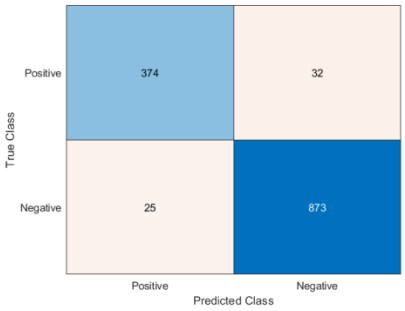

Fig. 9. Confusion Matrix between the Predicted Class and the True Class.

The SMO-based SVM model provides the following confusion matrix on Opinion Data's test data from the University of Illinois, Chicago [29]. The confusion matrix among the predicted class and true class is given in Fig. 9.

VII. Conclusion

The increasing corpus of text data brings challenges to the data storage and the text analytics' computational effort. These papers propose a framework for offering the Opinion Mining design process as a Cloud Services. The subscribers can avail themselves of fast and cost-effective services for the opinion analysis on their text corpus data. The paper proposes two distinguished methods as a Vedar based score computation and another as an SVM-based learning model as an opinion analytics engine. The score-based algorithm performs well on the small dataset, whereas the learning-based model is computationally effective on the large corpus.

The proposed algorithm for futuristic research can be considered with different datasets. Further, the given algorithm can be incorporated in other cloud services where opinion mining is necessary. Also, the security parameter can be incorporated in the ongoing and future researches in opinion mining.

References

[1] Durao, Frederico, Jose Fernando S. Carvalho, Anderson Fonseka, and Vinicius Cardoso Garcia. "A systematic review on cloud computing." The Journal of Supercomputing 68, no. 3 (2014): 1321-1346.

[2] Y. Hung, "Investigating How the Cloud Computing Transforms the Development of Industries," in IEEE Access, vol. 7, pp. 181505-181517, 2019, doi: 10.1109/ACCESS.2019.2958973.

[3] R. Neuhold, H. Gursch, R. Kern and M. Cik, "Driver's dashboard – using social media data as additional information for motorway operators," in IET Intelligent Transport Systems, vol. 12, no. 9, pp.

1116-1122, 11 2018. doi: 10.1049/iet-its.2018.5337

[4] S. Sakr, Z. Maamar, A. Awad, B. Benatallah, and W. M. P. Van Der Aalst, "Business Process Analytics and Big Data Systems: A Roadmap to Bridge the Gap," in IEEE Access, vol. 6, pp. 77308-77320, 2018. doi: 10.1109/ACCESS.2018.2881759

[5] S. Liu, Y. Chen, H. Wei, J. Yang, K. Zhou, and S. M. Drucker, "Exploring Topical Lead-Lag across Corpora," in IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 1, pp. 115-129, 1 Jan. 2015. doi: 10.1109/TKDE.2014.2324581

[6] A. Benito-Santos and R. Therón Sánchez, "Cross-Domain Visual Exploration of Academic Corpora via the Latent Meaning of User-Authored Keywords," IEEE Access, vol. 7, pp. 98144-98160, 2019. doi: 10.1109/ACCESS.2019.2929754

[7] R. C. Basole, H. Park and R. O. Chao, "Visual Analysis of Venture Similarity in Entrepreneurial Ecosystems," in IEEE Transactions on Engineering Management, vol. 66, no. 4, pp. 568-582, Nov. 2019. doi: 10.1109/TEM.2018.2855435

[8] Q. Han, D. Thom, M. John, S. Koch, F. Heimerl, and T. Ertl, "Visual Quality Guidance for Document Exploration with Focus+Context Techniques," in IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 8, pp. 2715-2731, 1 Aug. 2020.

[9] V. Raghupathi, Y. Zhou, and W. Raghupathi, "Legal Decision Support: Exploring Big Data Analytics Approach to Modeling Pharma Patent Validity Cases," in IEEE Access, vol. 6, pp. 41518-41528, 2018.doi: 10.1109/ACCESS.2018.2859052

[10] S. Kumar and M. Singh, "Big data analytics for the healthcare industry: impact, applications, and tools," in Big Data Mining and Analytics, vol. 2, no. 1, pp. 48-57, March 2019.doi: 10.26599/BDMA.2018.9020031

[11] Y. Keneshloo, T. Shi, N. Ramakrishnan, and C. K. Reddy, "Deep Reinforcement Learning for Sequence-to-Sequence Models," in IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 7, pp. 2469-2489, July 2020. doi: 10.1109/TNNLS.2019.2929141

[12] F. Amalina et al., "Blending Big Data Analytics: Review on Challenges and a Recent Study," in IEEE Access, vol. 8, pp. 3629-3645, 2020. doi: 10.1109/ACCESS.2019.2923270

[13] J. Koven, C. Felix, H. Siadati, M. Jakobsson and E. Bertini, "Lessons Learned Developing a Visual Analytics Solution for Investigative Analysis of Scamming Activities," in IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 1, pp. 225-234, Jan. 2019. doi: 10.1109/TVCG.2018.2865023

[14] M. El-Assady, R. Sevastjanova, F. Sperrle, D. Keim, and C. Collins, "Progressive Learning of Topic Modeling Parameters: A Visual Analytics Framework," in IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 382-391, Jan. 2018. doi: 10.1109/TVCG.2017.2745080

[15] W. Wang, K. Zhu, H. Wang, and Y. J. Wu, "The Impact of Sentiment Orientations on Successful Crowdfunding Campaigns through Text Analytics," in IET Software, vol. 11, no. 5, pp. 229-238, 10 2017. doi: 10.1049/iet-sen.2016.0295

[16] Y. Lu, H. Wang, S. Landis, and R. Maciejewski, "A Visual Analytics Framework for Identifying Topic Drivers in Media Events," in IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 9, pp. 2501-2515, 1 Sept. 2018. doi: 10.1109/TVCG.2017.2752166

[17] S. Gad et al., "ThemeDelta: Dynamic Segmentations over Temporal Topic Models," in IEEE Transactions on Visualization and Computer Graphics, vol. 21, no. 5, pp. 672-685, 1 May 2015. doi: 10.1109/TVCG.2014.2388208

[18] K. Kurzhals, M. John, F. Heimerl, P. Kuznecov and D. Weiskopf, "Visual Movie Analytics," in IEEE Transactions on Multimedia, vol. 18, no. 11, pp. 2149-2160, Nov. 2016. doi: 10.1109/TMM.2016.2614184

[19] A. Ojo and N. Rizun, "Enabling Deeper Linguistic-Based Text Analytics—Construct Development for the Criticality of Negative Service Experience," in IEEE Access, vol. 7, pp. 169217-169256, 2019. doi: 10.1109/ACCESS.2019.2947593

[20] R. Karam, R. Puri, and S. Bhunia, "Energy-Efficient Adaptive Hardware Accelerator for Text Mining Application Kernels," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 24, no. 12, pp. 3526-3537, Dec. 2016. doi: 10.1109/TVLSI.2016.2555984

[21] R. Vatrapu, R. R. Mukkamala, A. Hussain and B. Flesch, "Social Set Analysis: A Set Theoretical Approach to Big Data Analytics," in IEEE Access, vol. 4, pp. 2542-2571, 2016. doi: 10.1109/ACCESS.2016.2559584

[22] S. Liu, J. Yin, X. Wang, W. Cui, K. Cao, and J. Pei, "Online Visual Analytics of Text Streams," in IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 11, pp. 2451-2466, 1 Nov. 2016. doi: 10.1109/TVCG.2015.2509990

[23] M. Sun, P. Mi, C. North, and N. Ramakrishnan, "BiSet: Semantic Edge Bundling with Biclusters for Sensemaking," in IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 1, pp. 310-319, 31 Jan. 2016.doi: 10.1109/TVCG.2015.2467813

[24] I. Cho, W. Dou, D. X. Wang, E. Sauda, and W. Ribarsky, "VAiRoma: A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History," in IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 1, pp. 210-219, 31 Jan. 2016. doi: 10.1109/TVCG.2015.2467971

[25] S. Liu et al., "Bridging Text Visualization and Mining: A Task-Driven Survey," in IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 7, pp. 2482-2504, 1 July 2019. doi: 10.1109/TVCG.2018.2834341

[26] H. Strobelt, D. Oelke, B. C. Kwon, T. Schreck, and H. Pfister, "Guidelines for Effective Usage of Text Highlighting Techniques," in IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 1, pp. 489-498, 31 Jan. 2016. doi: 10.1109/TVCG.2015.2467759

[27] D. Park, S. Kim, J. Lee, J. Choo, N. Diakopoulos and N. Elmqvist, "ConceptVector: Text Visual Analytics via Interactive Lexicon Building Using Word Embedding," in IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 361-370, Jan. 2018. doi: 10.1109/TVCG.2017.2744478

[28] “Fasttext”, https://fasttext.cc/docs/en/english-vectors.html, Retrived on 26-02-2021

[29] “CS” https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html, Retrived on 26-02-2021

[30] X. Tian and K. Benkrid, "Mersenne Twister Random Number Generation on FPGA, CPU, and GPU," 2009 NASA/ESA Conference on Adaptive Hardware and Systems, San Francisco, CA, 2009, pp. 460-464, doi: 10.1109/AHS.2009.11.

[31] Gilbert CH, Hutto E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. InEighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. Social. gatech. edu/papers/icwsm14. vader. hutto. pdf 2014 Jun (Vol. 81, p. 82).

[32] Hu, M., and Liu, B., 2004, July. Mining opinion features in customer reviews. In AAAI (Vol. 4, No. 4, pp. 755-760).

[33] Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A. Advances in pre-training distributed word representations. arXiv preprint arXiv:1712.09405. 2017 Dec 26.

[34] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word Cloud Explorer: Text Analytics Based on Word Clouds," 2014 47th Hawaii International Conference on System Sciences, Waikoloa, HI, 2014, pp. 1833-1842, doi: 10.1109/HICSS.2014.231.

[35] Risch J, Kao A, Poteet SR, Wu YJ. Text visualization for visual text analytics. InVisual data mining, 2008 (pp. 154-171). Springer, Berlin, Heidelberg.

[36] Zixuan Ke, Bing Liu, Hao Wang, and Lei Shu. Continual Learning with Knowledge Transfer for Sentiment Classification. to appear Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD-2020), Ghent, Belgium, 14-18, September 2020.