Analyzing Behavior of Cancer Patients using Machine Learning Techniques

Jaswinder Singh, Sandeep Sharma

Abstract*: The online discussion forums and blogs are very vibrant platforms for cancer patients to express their views in the form of stories. These stories sometimes become a source of inspiration for some patients who are anxious in searching the similar cases. This paper proposes a method using natural language processing and machine learning to analyze unstructured texts accumulated from patient's reviews and stories. The proposed methodology aims to identify behavior, emotions, side-effects, decisions and demographics associated with the cancer victims. The pre-processing phase of our work involves extraction of web text followed by text-cleaning where some special characters and symbols are omitted, and finally tagging the texts using NLTK's (Natural Language Toolkit) POS (Parts of Speech) Tagger. The post-processing phase performs training of seven machine learning classifiers (refer Table 6). The Decision Tree classifier shows the higher precision (0.83) among the other classifiers while, the Area under the operating Characteristics (AUC) for Support Vector Machine (SVM) classifier is highest (0.98).

Index Terms*: Cancer Patients, Decision Tree, Feature Extraction, Machine Learning, Natural Language Processing.

I. INTRODUCTION

The popularity of the internet in the last few years has witnessed a significant growth in the number of internet users who seek medical information from online sources via social media information from online sources via social media platforms. These platforms make groups of users of users depending upon their personal choice or medical information they searched for. The discussions over these groups are well organized, where each thread of discussion commenced with a query and followed by responses in the form of personal experiences of patients and their health concerns(Ni, Xiao, Zhong, & Feng, 2018). These discussion threads are rich sources of clinical information. The recent researchers show that this valuable information can be harnessed from social media and exploited for various aspects like decision making, emotion mining, and an aid to recovery process of patients. For example, women affected with breast cancer share their stories online on social media groups without any hesitations(Gooden & Winefield, 2007). The informal talks sometimes results into informative and healthy discussions. The shared electronic-health-records may be valuable of a researcher or physician while doing some case-study. These electronic-health-records contain disease information, its symptoms, complications and severity. However, the records are well documented and narrative, but researchers still need to review them manually to identify the required information. The Natural Language Processing (NLP) provides a wide variety of tools for text processing. Retrieving key-words from electronic records and prioritization followed by indexing make a great deal of time management by expediting the preprocessing of data. The application of NLP and machine learning techniques facilitates many applications ranging from diagnosis, classification to adverse-drug-reactions. The diagnosis, classification involves identification of symptoms from textual reports and extending the secondary analysis from disease detection to sub-type evaluation. Another aspect of NLP application may find novel phenotypes(Zeng, Deng, Li, Naumann, & Luo, 2019). The conventional method for new disease discovery involves the adherence to the set of pre-defined criteria made by domain experts. On the other hand, the machine learning and NLP based methods detect hidden data coherence from social media and electronic records, this may promotes discovery of new phenotypes. Furthermore, converting unstructured text into structured clinical data has benefitted trial recruitment. The increased volumes of clinical data have attracted computational systems for its evaluation and identification of diseases. Also the interaction between different genes gains attraction in the upcoming years. The mining of pharmaco-genomics demands automated tools for phenol-typing, scanning, summarizing and retrieving the required phenotypes(Nikfarjam, Sarker, O'Connor, Ginn, & Gonzalez, 2015). The researchers these days are experimenting with semantics and syntactic analysis using different machine learning constructs for mining gene relations from these clinical records. The interaction between two drugs happens when an activity of one drug affects the event initiated by another drug. Such information is very crucial for a physician before prescribing personalized drugs to an infected person. Drug-Drug-Interactions (DDI) from clinical notes and patients personal history can be manually located, but this task requires huge efforts and hence consumes critical times of a patient(Zeng et al., 2019). The machine learning tools based on NLP can parse voluminous clinical records in a great deal of accuracy and processing time. Another example concerned with lung cancer can be observed through social media posts. Beside genetic issues, lung cancer depends upon various lifestyle factors. Textual description over social media about adverse effects of lung cancer shows that most of the cases fell into tobacco consumption. The second most prominent factor after the tobacco is air pollution. However, the diet, aging and hierarchical history of a patient also counts into

Revised Manuscript Received on July 05, 2019.

Jaswinder Singh, Department of Computer Science, Guru Nanak Dev University, Amritsar, Punjab, India.

Sandeep Sharma, Department of Computer Science & Engineering, Guru Nanak Dev University, Amritsar, Punjab, India.

mortality due to lung cancer(De Silva et al., 2018). Several other factors affect the human behavior, including socioeconomic level, medical fitness level, after birth expansion in the lungs, demographic structure and gender of a patient to name a few. The social media data combined with clinical data can be used for fact finding. The prediction of survival time in case of breast cancer can be evaluated through various factors viz., tumor size, stage of tumor, diagnosis-age, patient's cancer history, hierarchical data and immunity of patient to name a few(Mihaylov, Nisheva, & Vassilev, 2019). The machine learning techniques are trained using normalized parameters and the primary aim is to predict the survival time based on pre-defined features. As a benchmark American Cancer Society suggests average 83% survival for a decade. US government is working in this area to improve survival time by providing sustainable health facilities.

II. RELATED WORK

(De Silva et al., 2018) proposed an AI based analytical framework called PRIME (Patient Reported Information Multidimensional Exploration) in order to support the social media empowered patients. The proposed task suggested the importance of machine learning (ML) through OSG (online support group) for cancer care as well as for cancer treatment decisions. They demonstrated the ML and NLP (natural language processing) techniques on unstructured text discussions accrued in OSG and intelligently extracted the patient reported demographics, expressions of emotions, pre/post decisions making as well as Positive/Negative patient behaviors. They synthesized and suggested that the extracted information is helpful in pre/post decision making on personalized health care as well as for the expansion of health care policy guidelines. (Zeng et al., 2018) proposed the model which identifies breast cancer local recurrence through Electronic Health Records (EHRs) with the help of machine learning (ML) and natural language processing (NLP) techniques. Specially, the authors followed the SVM (support vector machine) with the NLP technique and achieved the AUC (Area under curve) 0.93 during cross validation and the 0.87 during the testing process in contrast to labor/manual intensive chart review, which is a time consuming process. They compared their model with 3 classifiers using either filtered metamap, full metamap concepts or bag of words. The authors suggested the importance of ML for cancer (wide and versatile area of research) specially, breast cancer. (Zeng et al., 2019) in their other paper reviewed the applications of different natural language processing (NLP) methods for Electronic Health Records (EHRs) based on the computational phenotyping. The phenotypes are observed by experts based on the bio-chemical processes or behaviors, physical appearance, e.g. BMI, height, weight, etc. The authors elaborated the rule based systems and analyzed the significance of supervised and unsupervised learning along with the deep learning. They revealed the challenges and opportunities for NLP based computational phenotyping for better future models, interoperability and their generalization.

(Luo, Riedlinger, & Szolovits, 2014) proposed the methodology for cancer gene and its pathway prioritization using text mining and suggested the importance of the pathway prioritization over the gene prioritization. They analyzed the heterogeneous data sources or we can say huge biomedical loads along with the challenges of the text mining in gene prioritization. The authors elaborate the different techniques of text mining for prioritization and suggested the future directions over the gene prioritization such as text network translation and pathway prioritization for finer semantic information. (Ni et al., 2018) proposed the methodology to predict the lung cancer mortality with the help of support vector machine (SVM) through multiple human behavior indicators. The authors suggested that In-spite of cancer sample size, there are multiple human behavior indicators which are very helpful in detection of lung cancer. In their study, the authors took 30 human behavior indicators of 7 categories such as alcohol consumption and tobacco smoking, air pollution, demographic structure, socioeconomic status, medical level, working culture and food structure. They have taken data set of 13 countries from the WHO (world health organization) and organization of economic corporation and development (OECD) database. The dataset belongs to the European and American countries and one of the important findings of the authors is that specially, Eastern Europe people having more lung cancer due to shortage of time for exercising. Their novel approach shows its importance over the traditional prediction approaches based on the clinical data sets. (Mihaylov et al., 2019) proposed the methodology for survival prognosis in breast cancer studies using machine learning (ML). Basically, they predict the survival time based on the novel feature called tumor integrated clinical feature (TICF), which combines the tumor stage, size and age at the diagnosis. They took two heterogeneous data sets from the Cancer Genome Atlas, the first dataset of 498 patients having clinical information and the second dataset of 2000 patients having genomic profile data as well as clinical information. They did the preprocessing of the data set through a Python scikit-learn library for normalizing the datasets in order to propose the tumor integrated clinical feature. The authors deployed the different ML approaches such as Lasso regression, linear support vector regression, k-neighborhood regression, kernel ridge regression and decision tree regression. All the ML techniques generating the promising results, but the kernel ridge regression was a producing better result over the other machine learning techniques. (T. Bandaragoda et al., 2018; De Silva et al., 2018) suggested the need of inclusive environment to exchange the information and proposed the novel way to extract the personalized knowledge from the online support group by performing the text mining. The author performed the experiment on the online active support group containing 8000000 posts and the 72066 active members. They performed their experiment using WEKA and deployed the naïve based and random forest machine learning classifier and achieved the f-measures as 0.80 and 0.85, respectively. The experiment performed on the dataset accentuates the importance of the proposed approach and unlock the extensive spectrum of personalized knowledge hidden within the

crowd-sourced contents. (T. R. Bandaragoda, De Silva, Alahakoon, Ranasinghe, & Bolton, 2018) in their other paper proposed the PRIME (Patient Reported Information Multidimensional Exploration) framework in order to investigate emotions as well as other factors of the PC (prostate cancer) patients with low intermediate risk through web based cancer support group discussions. In their study the authors analyzed the QOL (Quality of Life) emotions and the side effects of 6084 patients those who are undergoing EBRT (external beam radiotherapy) RP (radical prostatectomy) and AS (active surveillance). The authors identifies that the patients having age less than forty expressed more positive and negative emotions as compared to other age group patients together with that they also analyzed the behavior of partners of these patients and revealed that they expressed more negative emotions as compared to patients. The author proposed novel ensemble machine and deep learning algorithmic approach based on the clustering, classification association rules and NLP on crowd-sourced contents of discussion board. (Ruthven, Buchanan, & Jardine, 2018) analyzed the emotions of the young first time mother those who are looking some information and support on the discussion boards. The authors studied the isolated, worried and overwhelmed young mother's posts through the LIWC (Linguistic Inquiry and Word Count) dictionary-based software and deployed the statistical techniques to reveal the results. The authors took the dataset from the NetMums' "Young Parents Support" forum and the BabyCentre forum. They analyzed the 174 posts of the 162 participants and performed the textual analyses and classified the emotions into three categories such as preoccupation emotions, response emotions, and interaction emotions after that they revealed that many requests for information by youthful first-time mothers are aggravated by negative-emotions. (Gooden & Winefield, 2007) did the thematic analysis of breast cancer (BC) and prostate cancer (PC) about gender difference and similarities held online on the discussion boards. The author did the analysis using quasi-numerative approach and mixed methodology of both grounded theory. They revealed how men and women both share their emotions and knowledge over the discussion board with their fellow sufferers. The authors also suggested the importance and need of discussion boards for the BC and PC patients and expresses their views how discussion boards can improve the quality of life of the sufferers. (M. Y. Kim, Xu, Zaiane, & Goebel, 2013) proposed the methodology to extract the patient information present in the noisy Tele-Health texts. They performed their experiment using unsupervised machine learning and NLP approach together with Damerau-Levenshtein distance. After removing the noise from the data they performed their experiment over the normalized patient information. The authors have performed the experiment over the 3328 sentences of 200 patients and significantly achieved the 77.94% F-measure and their method outperformed the MetaMap by 17%. Their results show that they have achieved reasonable performance with their proposed method.

Table 1: Summary of Literature Survey

Researcher	Data Sets	Techniques	Results and Discussions
(De Silva et al., 2018)	Cancer online discussion board	Unstructured Machine Learning (ML) & Natural Language Processing (NLP)	Proposed PRIME framework, helpful in cancer treatment decisions.
(Zeng et al., 2018)	Breast cancer Electronic Health Records (EHRs)	SVM (support vector machine) and NLP	Identifies breast cancer local recurrence and achieved AUC as 0.93 and cross validation as 0.87
(Zeng et al., 2018)	Cancer Electronic Health Records (EHRs)	Rule based supervised and unsupervised learning along with the deep learning as well as NLP	Revealed challenges and opportunities for NLP based computational phenotyping for better future models
(Luo et al., 2014)	Heterogeneous data sources of cancer gene	Text mining through NLP	Pathway prioritization of genes rather than gene prioritization through NLP for finer semantic information
(Ni et al., 2018)	WHO (world health organization) and organization of economic corporation and development (OECD) database	SVM	Their novel approach shows its importance over the traditional prediction approaches based on the clinical data sets.
(Mihaylov et al., 2019)	Cancer Genome Atlas	ML techniques implementation using Python scikit-learn library	Predicted the survival time based on the novel feature called tumor integrated clinical feature (TICF) through kernel ridge regression ML approach.

(T. R. Bandaragoda et al., 2018)	Online active support group containing 8000000 posts and the 72066 active members	ML implementation using WEKA	Extracted the personalized knowledge from the online support group by performing the text mining and achieved f-measure 0.85.
(T. Bandaragoda et al., 2018)	Online Discussion group of 6084 patients those who are undergoing EBRT (external beam radiotherapy) RP (radical prostatectomy) and AS (active surveillance).	ML, Deep Learning and NLP	The author proposed novel ensemble machine and deep learning algorithmic approach based on the clustering, classification association rules and NLP on crowd-sourced contents of discussion board
(Ruthven et al., 2018)	NetMums' "Young Parents Support" forum and the BabyCentre forum	LIWC (Linguistic Inquiry and Word Count) dictionary-based software	The Authors analyzed the emotions of the young first time mother those who are looking some information and support on the discussion boards.
(Gooden & Winefield, 2007)	Breast Cancer (BC) and Prostate Cancer (PC) online discussion Borad	Quasi-numerative and mixed methodology of both grounded theory	The authors did the thematic analysis of breast cancer (BC) and prostate cancer (PC) about gender difference and similarities held online on the discussion boards. Which improves QOL (Quality of Life)
(M. Y. Kim et al., 2013)	Noisy Tele-Health texts of Patients	Unsupervised ML and NLP approach together with Damerau-Levenshtein distance method	The authors have performed the experiment over the 3328 sentences of 200 patients and significantly achieved the 77.94% F-measure and their method outperformed the MetaMap by 17%.

III. PROPOSED METHODOLOGY

The patient stories and reviews from six different sources (Refer Table 2) are extracted using bS4 library of Python 3.6. the data scrape operation for collecting around 1200 stories took around one hour on core i3 machine with 8 GB RAM. The extracted data is then exported into CSV file followed by conversion into Unicode format in order to get rid of special characters from web-text. The preprocessing of textual data includes data-cleaning operations where removal of hyperlinks, hash characters and non-English words are processed using NTLK library. The train-test split module converts CSV into two files (80% for training and 20% of testing and validation). The major part of the dataset undergoes tagging using standard NLTK's tagger followed by extraction of filter-terms (using POS Tagger). The training of seven machine learning modules is preceded using training set and followed by prediction of patient's behavior on testing set. The performance after 10-fold cross validation is averaged (Refer Table 5 and Figure 1).

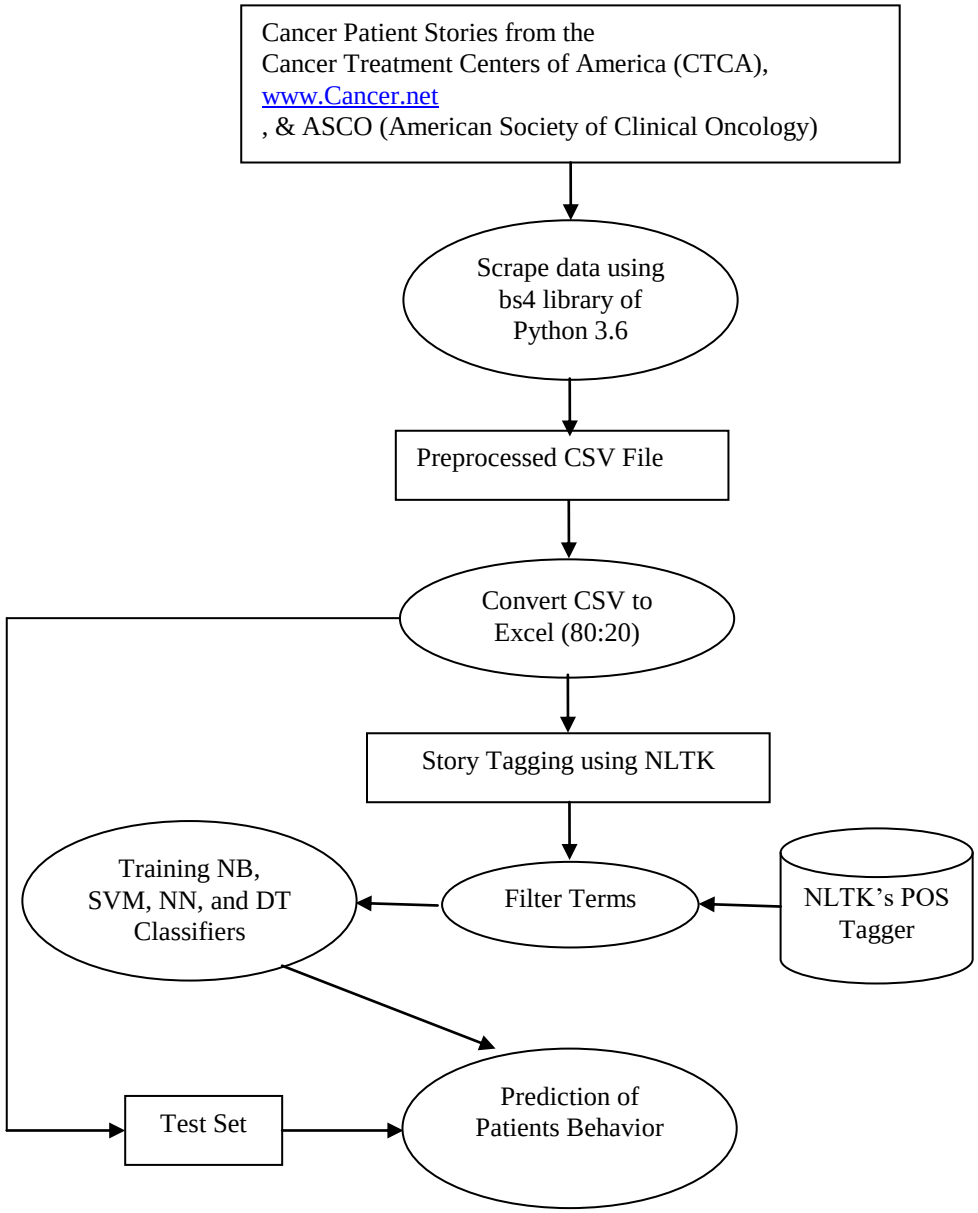

Figure 1: Workflow of proposed system for Behavior Evaluation

A Feature Extraction

The Figure 2 shows top-ranked terms extracted using feature tool of sklearn. It finds the frequency of each term (word or phrase) using word-matrices and reports most frequent terms based on threshold frequency set by user. Here, our frequency was fifteen; it has reported eleven concepts which occurred more than 150 times in patient's reviews.

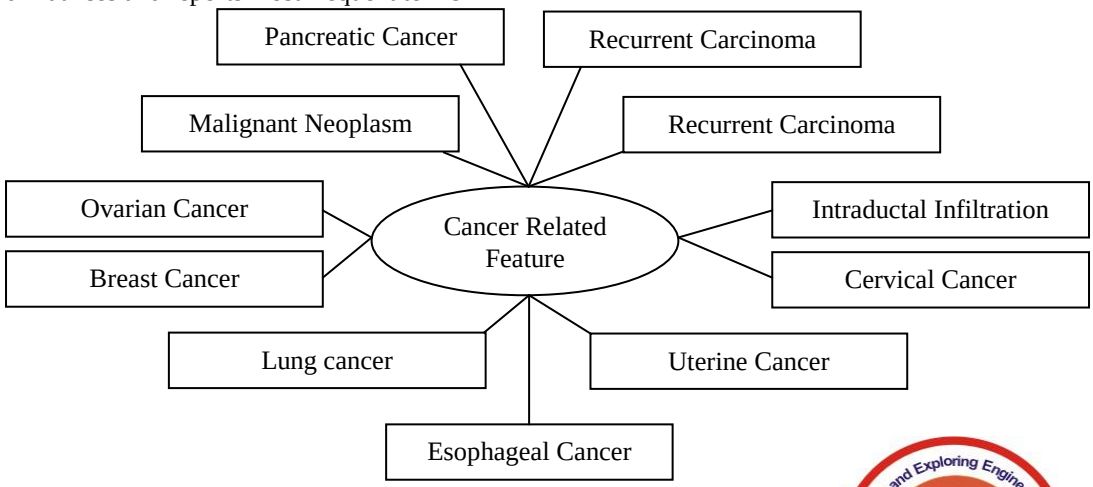

Figure 2: Feature tool Python 3.6 for obtaining Top- Ranked Concepts

B Outcome of Feature Extraction

The Figure 3 highlights ten recovery tips and traits extracted through keyword-search tool of Python's NLTK. These recovery traits are the recommendation of our proposed system which can be helpful for readers and bloggers in their recovery process.

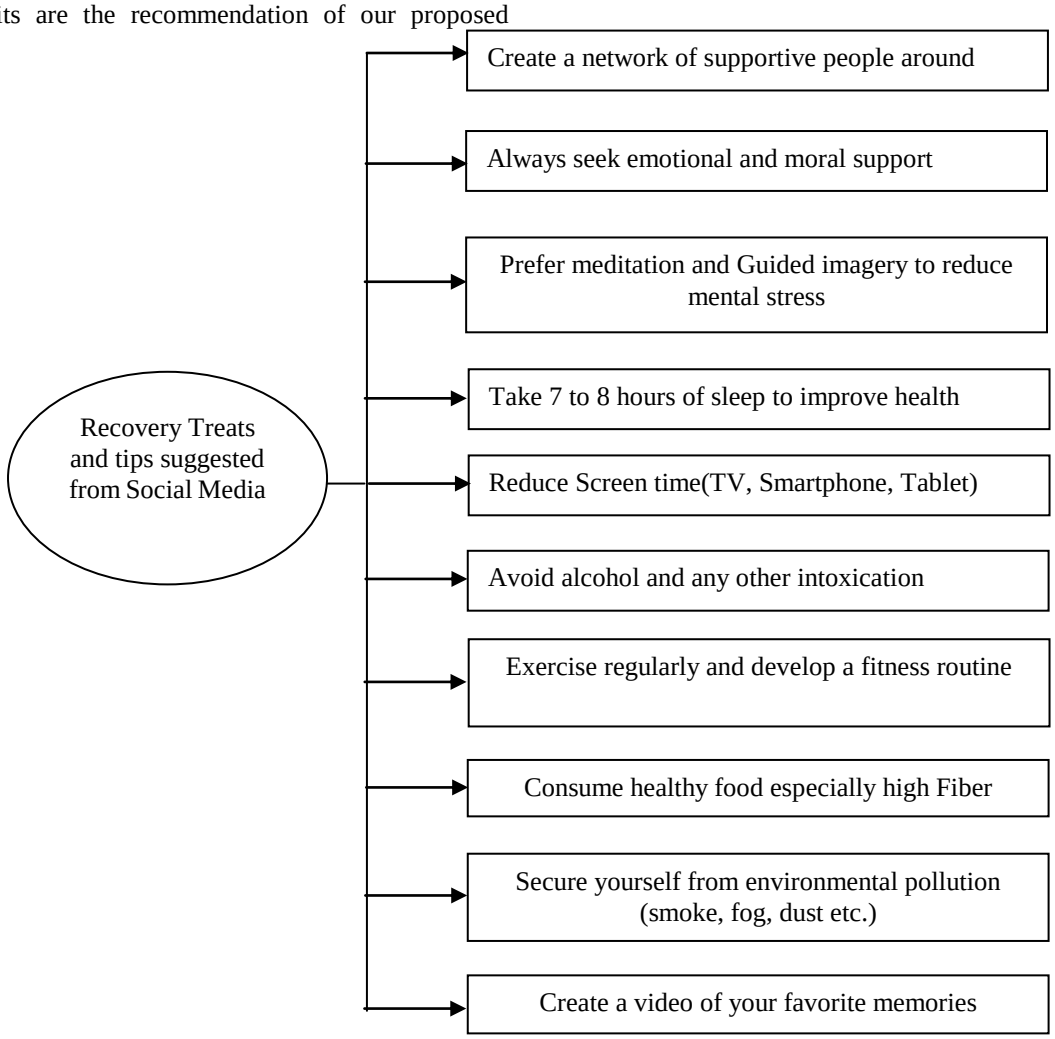

Figure 3: Recovery Traits obtained through keyword search tool for NLTK

3.3 Statistics of Dataset

The Table 2 shows horrifying statistics for cancer mortality in USA from last decade. These statistics are taken from CTCA's website. Here, LC denotes Lung Cancer, BC as Breast Cancer, OC as Ovarian Cancer, PC as Pancreatic Cancer, SC as Skin Cancer, and UC is Uterine Cancer. This table clearly shows that numbers of cases in BC are highest among other cancers, and this trend remains in the existence from last 9 years. The second largest trend is for UC. These two major cancer categories are largely discussed over web forums and discussions blogs.

Table 2: Statistics from Cancer Treatment Center of America

Year	No. of Deaths					
	LC	BC	OC	PC	SC	UC
2011	60	1035	150	15	652	903
2012	58	1010	148	11	598	854
2013	53	1030	142	3	633	784
2014	52	1008	155	10	640	838
2015	55	946	143	13	578	837
2016	56	856	161	8	570	872
2017	58	1030	132	3	533	730
2018	50	850	121	9	547	712

2019	32	412	56	1	138	216

The large numbers of discussions are going on where a patient expresses their reviews on blog and forums. Since, our work considers only cancer patients, we have taken only 1262 patient reviews from six different sources mentioned in the Table 3 of the manuscript.

D Data Sources

The proposed methodology for behavior prediction for cancer patients is applied with the aim of data classification. Our proposed methodology extracts emotion words from patient reviews and stories and classifies them into positive or negative. The training of seven machine learning methods is done using six data sources as mentioned in Table 3. The power of these methods in improving the accuracy of classification depends upon the type of dataset. Our datasets are confined to patient reviews and stories mentioned on blogs and forums.

Table 3: Data Sources

Datasets Sources	URL	No. of Reviews/ Stories
Cancer Treatment Centers of America (CTCA)	www.cancercenter.com	484
Cancer .Net	www.cancer.net/blog/	315
American Society of Clinical Oncology	www.asco.org	218
Healing well community	www.healingwell.com	152
The cancer Forums	www.cancerforums.net/blog.php	78
National Cancer Institute USA	www.cancer.org	15

IV. RESULTS AND DISCUSSIONS

The Table 4 shows six example texts from patient reviews (these for each positive-negative). The extracted sentences contain emotional words found by our proposed method.

Table 4: Extracted emotional words from patient reviews

Patient's Reviews (Extracted emotional sentences)	Emotion words & Phrases
My cancer journey brought several ups and downs. I had lived good and bad days. Sometimes, I found myself helpless and trapped, whereas some days were positive like "I can make it", "I can help everything" (posted by Elaine B. on Feb. 18, 2018 at CTCA)	Ups, downs, good and bad days, helpless, trapped, I can make it, I can help everything
I had got used to the ordeal, great fear of recurrence, the complex treatments, and waves of hope and despair. I was given a strong antibiotic, which made me weak and more infected. Each day was a struggle for survival. (Posted by Christine B. on Mar 04, 2018 at CTCA)	Ordeal. Great fear of Recurrence, Waves of hope, Despair, Weak, Struggle for survival
When I received the diagnosis, I was shocked and overwhelmed. I was known that outcome of any complex treatment is uncertain. I felt lonely. At one point of time, I am a mother of three daughters and at other point of time I am a cancer patient. I had no idea that what will happen to me. (Posted by Jannifer on Mar 04, 2018cat CTCA)	Shocked, Overwhelmed, Complex treatment is uncertain, lonely, cancer patient
The Quality of care I received at CTCA was excellent. All the time I found somebody around me to answer my queries. The doctors, nurses, and staff were beyond my expectations. I am grateful for their love and support. (Posted	Excellent, Care, Beyond my expectations, Grateful, Love and Support

by Julie Ulva on Mar 1, 2018 at CTCA)	
The overall treatment at CTCA was very convenient to me. Everything was arranged under one roof. The wait time is always minimal. Their dealings helped me to relax and keep focusing on healing. My caregivers were well behaved and looked forward to every of my demands. My friends and staff celebrated my successful treatment and arranged a treat.	Convenient, Relax, Healing, Well behaved, Celebrated, Successful treatment.
I was not completely surprised at the diagnosis because of my mother's history. I was not even scared, but I was concerned that how users I going to get through this. It was disheartening that I was having Lymph edema, radiation, and chemotherapy. Due to pain, I could not even fold the laundry. (Posted by Kirenn N. on Apr 22, 2018 at CTCA)	Not completely, Surprised, Not even scared, Disheartening, Pain

A Extracted Emotion Words

The Table 5 of the manuscript presents range of emotion words taken from dataset. The word similarities are evaluated using NLTK's word-net corpus(Y. Kim, 2014). The synsets retrieved in two different lists are maintained and tagged positive or negative in the pre-processing phase of proposed methodology (Khan, Qamar, & Bashir, 2016). The maximum value of similarity is taken from wup-similarity method of word-net. The value above 0.9 confirms that the words are of same category(Mikolov, Corrado, Chen, & Dean, 2013). The classifier later on considers this value for classification of newly retrieved words from dataset.

Table 5: Emotion words Extracted by proposed methodology

Positive	Negative
Happy, Terrific, Fab, Easy, Good, supportive, loving, cuddly treatment, care, caring, cynical, cooperative, cautious, friendly, amiable, resilient, decisive, intuitive, tenacious, proactive, persistent.	Sad, panic, deprived, fatigued, weak, dull, miserable, rude, numb, emotionally, drained, perplexed, anxiety, cramps, hopeless, defeated, killing cancer, blunt, hurtful, cry, teary, agitated.

B. Performance of Machine Learning Classifiers

The Table 6 illustrates the performance of seven machine learning methods in terms of Precision, Recall, F-measure and Area under operating characteristics. All results are based on pre-processed patient's reviews from dataset. The high precision of decision tree classifier among other models makes it appropriate for behavior evaluation. However, the AUC for SVM (Linear SVC) is highest (0.98) among the seven classifiers. The performance of SVM and Logistic Regression are similar in terms of precision for classification.

Table 6: Training Accuracies of Different Machine Learning Methods

Method	Precision	Recall	F-Measure	AUC
Naïve Bayes Gaussian	0.685	0.4525	0.60	0.723
Support Vector Machine (SVC)	0.775	0.60	0.63	0.785
Logistic Regression	0.79	0.55	0.70	0.81
Decision Tree	**0.83**	0.73	0.79	0.88
Neural Networks	0.69	0.48	0.58	0.79
Multinomial Naïve Bayes	0.695	0.4575	0.61	0.72
SVM (Linear SVC)	0.78	0.58	0.6325	**0.98**

C Snapshot of Python 3.6.0 Shell

The snapshot of the results performed using sklearn library of Python 3.6 is presented in the Figure 4, which depicts the performances of different classifiers as mentioned in the Table 6.

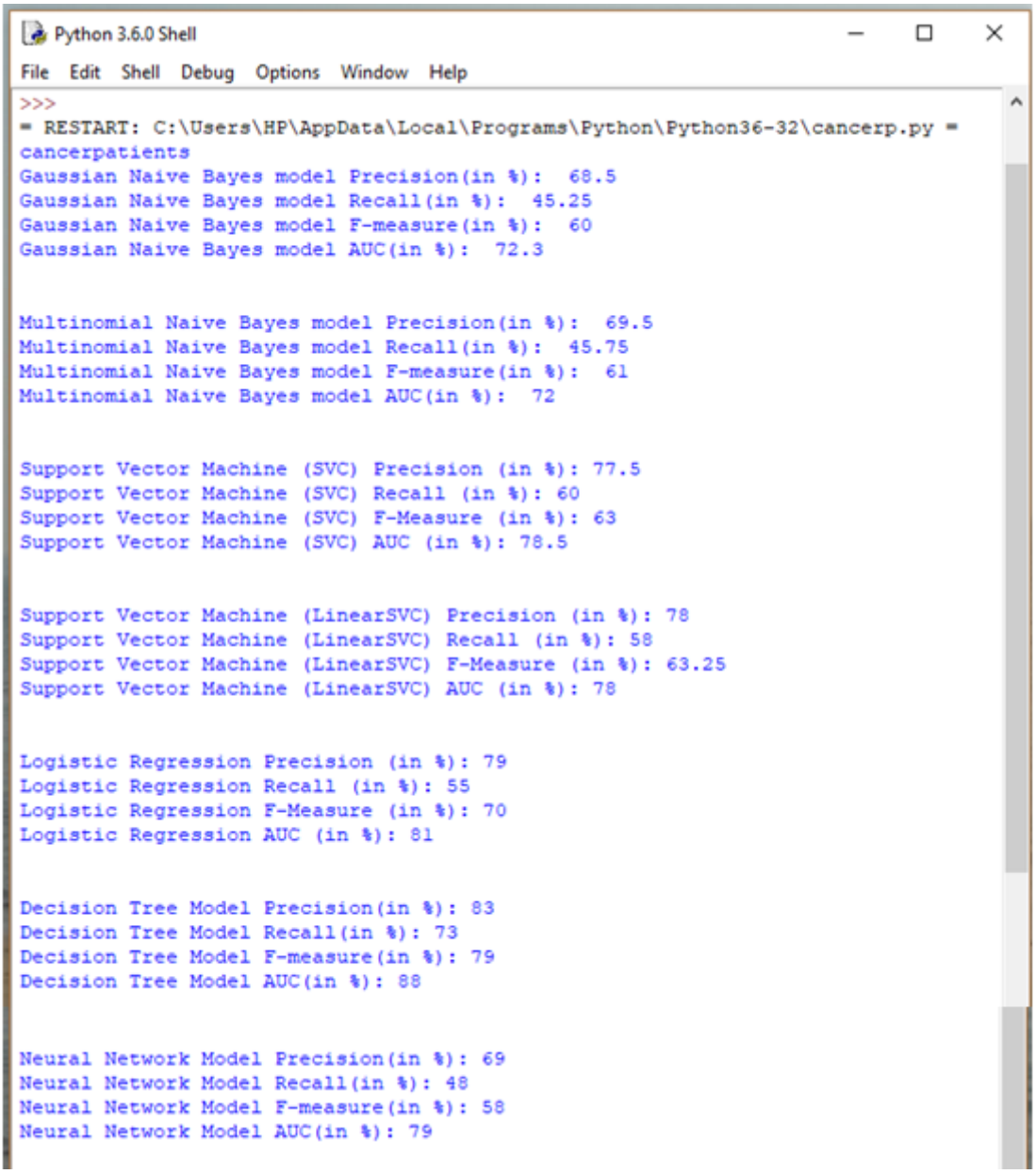

Figure 4: Snapshot of Results

V. CONCLUSION

The high performance of classifiers reveals that the proposed framework has a potential in for serving the patients on social media. The training precision is promising for decision tree classifier, which renders the classification task to extract behavioral and emotional aspects of patients for taking appropriate decisions in their treatment. The NLP based pre-processing of texts provide a way of aggregation and consolidation, which further helps patients in their decision making and behavior analysis. We have provided a framework using NLP and machine learning based techniques for the healthcare monitoring of cancer patients. The future scope of this work aims at the subjectivity of the concept and considers the further fine details from web texts in order to identify the true emotions of cancer patients.

REFERENCES

1. Bandaragoda, T. R., De Silva, D., Alahakoon, D., Ranasinghe, W., & Bolton, D. (2018). Text Mining for Personalized Knowledge Extraction From Online Support Groups. *Journal of the Association for Information Science and Technology*, *69*(12), 1446–1459. https://doi.org/10.1002/asi.24063
2. Bandaragoda, T., Ranasinghe, W., Adikari, A., de Silva, D., Lawrentschuk, N., Alahakoon, D., … Bolton, D. (2018). The Patient-Reported Information Multidimensional Exploration (PRIME) Framework for Investigating Emotions and Other Factors of Prostate Cancer Patients with Low Intermediate Risk Based on Online Cancer Support Group Discussions. *Annals of Surgical Oncology*, *25*(6), 1737–1745. https://doi.org/10.1245/s10434-018-6372-2
3. De Silva, D., Persad, R., Lawrentschuk, N., Iddamalgoda, L., Bolton, D., Osipov, E., … Gray, R. (2018). Machine learning to support social media empowered patients in cancer care and cancer treatment decisions. *Plos One*, *13*(10), e0205855. https://doi.org/10.1371/journal.pone.0205855
4. Gooden, R. J., & Winefield, H. R. (2007). Breast and prostate cancer online discussion boards: A thematic analysis of gender differences and similarities. *Journal of Health Psychology*, *12*(1), 103–114. https://doi.org/10.1177/1359105307071744
5. Khan, F. H., Qamar, U., & Bashir, S. (2016). SentiMI: Introducing point-wise mutual information with SentiWordNet to improve

sentiment polarity detection. *Applied Soft Computing Journal*, *39*, 140–153. https://doi.org/10.1016/j.asoc.2015.11.016
6. Kim, M. Y., Xu, Y., Zaiane, O., & Goebel, R. (2013). Patient information extraction in noisy tele-health texts. *Proceedings - 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013*, 326–329. https://doi.org/10.1109/BIBM.2013.6732511
7. Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. https://doi.org/10.3115/v1/D14-1181
8. Luo, Y., Riedlinger, G., & Szolovits, P. (2014). Text Mining in Cancer Gene and Pathway Prioritization. *Cancer Informatics*, *13s1*, CIN.S13874. https://doi.org/10.4137/cin.s13874
9. Mihaylov, I., Nisheva, M., & Vassilev, D. (2019). Application of machine learning models for survival prognosis in breast cancer studies. *Information (Switzerland)*, *10*(3), 1–13. https://doi.org/10.3390/info10030093
10. Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, 1–12. https://doi.org/10.1162/153244303322533223
11. Ni, D., Xiao, Z., Zhong, B., & Feng, X. (2018). Multiple Human-Behaviour Indicators for Predicting Lung Cancer Mortality with Support Vector Machine. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-34945-z
12. Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., & Gonzalez, G. (2015). Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, *22*(3), 671–681. https://doi.org/10.1093/jamia/ocu041
13. Ruthven, I., Buchanan, S., & Jardine, C. (2018). Isolated, overwhelmed, and worried: Young first-time mothers asking for information and support online. *Journal of the Association for Information Science and Technology*, *69*(9), 1073–1083. https://doi.org/10.1002/asi.24037
14. Zeng, Z., Deng, Y., Li, X., Naumann, T., & Luo, Y. (2019). Natural Language Processing for EHR-Based Computational Phenotyping. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *16*(1), 139–153. https://doi.org/10.1109/TCBB.2018.2849968
15. Zeng, Z., Espino, S., Roy, A., Li, X., Khan, S. A., Clare, S. E., … Luo, Y. (2018). Using natural language processing and machine learning to identify breast cancer local recurrence. *BMC Bioinformatics*, *19*(Suppl 17). https://doi.org/10.1186/s12859-018-2466-x

AUTHORS PROFILE

Mr. Jaswinder Singh pursued Master of Computer Application from Himachal Pradesh University, Shimla. Currently, he is pursuing Ph.D. in Computer Science and Engineering from Guru Nanak Dev University, Amritsar. His research interest is Wireless Sensor Network, Machine Learning.

Dr. Sandeep Sharma has done his B.E in Computer Science and Engineering, M.E in Computer Science and Engineering and Ph.D. His area of interest is Big Data, Cloud Computing and Parallel Processing. Currently, he is Head and Professor at Department of Computer Engineering and Technology, Guru Nanak Dev University Amritsar