SHRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA, INDORE

SHRI VAISHNAV SCHOOL OF MANAGEMENT

NAV-PRABANDHAN 2023

Certificate of Participation

This is to certify that **Dr. Anupama Chaudhari** of **KCES's Institute of Management and Research, Jalgaon**

presented a paper entitled '**AWARENESS ABOUT GREEN PRACTICES AMONG THE STAFF MEMBERS OF HEI'S OF THE JALGAON CITY' in Nav-Prabandhan** held on September 22-23, 2023.

Dr. Swati Oza	**Dr. Santosh Dhar**	**Dr. Upinder Dhar**
Conference Coordinator	Conference Chairperson	Vice Chancellor

G H Raisoni Institute of Engineering and Business Management Jalgaon

RAISONI GROUP
a vision beyond

An **AUTONOMOUS** Institution Affiliated to KBCNMU Jalgaon, NAAC Accredited **'A'** Grade

INSTITUTION'S INNOVATION COUNCIL (Ministry of Education Initiative)

MAGIC BY CMIA
MAGIC
Marathwada Accelerator for Growth & Incubation Council

G H Raisoni Centenary Decade Celebration 1930-2030

Department of Management

Organized

3rd International Conference on

INNOVATION, TECHNOLOGY & ENTREPRENEURSHIP (ITE-2024)

12th - 13th January 2024

CERTIFICATE

This is to certify that Dr. Anupama Chaudhari has presented a research paper on "Emotional Intelligence And Work Life Balance: A Study Of Dual Career Couples" in the 3rd International Conference.

Dr. Jyoti Jakhete
Convener

Dr. Jitendrasinh Jamadar
Convener

Dr. Yogita Patil
Convener

Dr. Sanjay Shekhawat
Dean Academics, GHRIEBM,Jalgaon

Dr. Preeti Agarwal
Director, GHRIEBM,Jalgaon

G H Raisoni Institute of Engineering and Business Management Jalgaon

RAISONI GROUP
a vision beyond

G H RAISONI CENTENARY DECADE CELEBRATION 1930-2030

An **AUTONOMOUS** Institution Affiliated to KBCNMU Jalgaon, NAAC Accredited **'A'** Grade

INSTITUTION'S INNOVATION COUNCIL (Ministry of Education Initiative)

MAGIC BY CMIA
MAGIC
Marathwada Accelerator for Growth & Incubation Council

Department of Management

Organized

3rd International Conference on

INNOVATION, TECHNOLOGY & ENTREPRENEURSHIP (ITE-2024)

12th - 13th January 2024

CERTIFICATE

This is to certify that Dr. Mamta Dahad has presented a research paper on "Emotional Intelligence And Work Life Balance: A Study Of Dual Career Couples" in the 3rd International Conference.

Dr. Jyoti Jakhete	**Dr. Jitendrasinh Jamadar**	**Dr. Yogita Patil**	**Dr. Sanjay Shekhawat**	**Dr. Preeti Agarwal**
Convener	Convener	Convener	Dean Academics, GHRIEBM,Jalgaon	Director, GHRIEBM,Jalgaon

G H Raisoni Institute of Engineering and Business Management Jalgaon

RAISONI GROUP
a vision beyond

An **AUTONOMOUS** Institution Affiliated to KBCNMU Jalgaon, NAAC Accredited **'A'** Grade

INSTITUTION'S INNOVATION COUNCIL

MAGIC

Department of Management

Organized

3rd International Conference on

INNOVATION, TECHNOLOGY & ENTREPRENEURSHIP (ITE-2024)

12th - 13th January 2024

CERTIFICATE

This is to certify that Ms.Dhanashree Dipak Chaudhari has presented a research paper on "Perceived Organizational Support and Employee Engagement Among Bank Employees in a Digital Era" in the 3rd International Conference.

Dr. Jyoti Jakhete	**Dr. Jitendrasinh Jamadar**	**Dr. Yogita Patil**	**Dr. Sanjay Shekhawat**	**Dr. Preeti Agarwal**
Convener	Convener	Convener	Dean Academics, GHRIEBM,Jalgaon	Director, GHRIEBM,Jalgaon

Khandesh College Education Society's

INSTITUTE OF MANAGEMENT & RESEARCH, JALGAON

NAAC Accredited 'A' Grade (CGPA 3.17) | NBA Accredited MBA Program

in Association with KBC North Maharashtra University, Jalgaon

3rd INTERNATIONAL CONFERENCE

ON

INNOVATION, AUTOMATION AND FUTURE TRENDS IN BUSINESS

Certificate of Participation

This is to certify that ~~Mr.~~ /Ms. /~~Dr.~~ /~~Prof.~~ Dhanashree Dipak Chaudhari

of KCES's Institute of Management & Research, Jalgaon.

has participated / presented a research paper entitled Perceived Organizational Support & Employee Engagement : The key to Improving Employee Performance In the Private Banking Sector. in the 3rd International Conference On Innovation, Automation And Future Trends In Business on March 25th, 2023.

Mr. Anilkumar Marthi	**Dr. Parag Narkhede**	**Prof. Shilpa K. Bendale**
(Conference Secretary)	(Conference Convener)	(Conference Chair)
		Director, KCES's IMR Jalgaon

Khandesh College Education Society's

INSTITUTE OF MANAGEMENT & RESEARCH, JALGAON

NAAC Accredited 'A' Grade (CGPA 3.17) | NBA Accredited MBA Program

in Association with KBC North Maharashtra University, Jalgaon

3rd INTERNATIONAL CONFERENCE

ON

INNOVATION, AUTOMATION AND FUTURE TRENDS IN BUSINESS

Certificate of Participation

This is to certify that ~~Mr.~~ /Ms. /~~Dr.~~ /~~Prof.~~ Dipali Sadashiv Patil

of KCES's Institute of Management and Research, Jalgaon

has participated / presented a research paper entitled Strategies for Managing work-related stress through work-life Balance theories. in the 3rd International Conference On Innovation, Automation And Future Trends In Business on March 25th, 2023.

Mr. Anilkumar Marthi	**Dr. Parag Narkhede**	**Prof. Shilpa K. Bendale**
(Conference Secretary)	(Conference Convener)	(Conference Chair)
		Director, KCES's IMR Jalgaon

Development of Code-Mixed Marathi-English Dataset for Hate Speech Detection

Prasad Joshi
JET's Zulal Bhilajirao Patil College, Dhule(M.S.)
K B C North Maharashtra University, Jalgaon(M.S.)
Dhule, India
sayprasadajoshi@gmail.com,
0000-0001-5522-6187

Varsha Pathak
Institute of Management and Research, Jalgaon(M.S.)
K B C North Maharashtra University, Jalgaon(M.S.)
Jalgaon, india
pathak.vmpathak.varsha@gmail.com

***Abstract*—In India, people commonly use a mix of English and their regional languages on social media, resulting in a substantial amount of code-mixed content. As of the present date, there is only a limited number of hate speech code-mixed datasets available for Indian languages, including Hindi-English, Bengali-Hindi, Malayalam-English, and Tamil-English. However, there are no resources or datasets available for Marathi-English code-mixed data with hate speech labels. This paper introduces a novel gold standard corpus for sentiment analysis of code-mixed text in Marathi-English, annotated by voluntary annotators. The gold standard corpus achieved an impressive Krippendorff's alpha score of above 0.9 for the dataset. Utilizing this new corpus, the paper establishes a benchmark for sentiment analysis in Marathi-English code-mixed texts.**

***Index Terms*—code-mixed, Marathi, hate, sentiment, dataset, machine learning**

I. Introduction

Hate speech is any method of communication, whether spoken or written against an individual or a group of people. These days, different social media platforms play a critical role in creating and spreading hate speech. Some of the recent incidences in India caused because of social media hate speech are - In July 2022, Maharashtra police filed over 600 cases against social media users for communal hateful content [1]. Due to fake news on social media, 33 people were killed in 69 mob violence cases related to rumours of child lifting [2]. A hateful Facebook post concerning the Prophet Muhammad triggered violent clashes in Bengaluru in 2020 [3]. A survey conducted by the Mumbai-based Association of Adolescent and Child Care in India (AACCI) found that schools in Mumbai and Gurgaon have risen in aggression due to social media [4]. Hence, such hate content is a crucial task in front of the research community. These social media comments are in monolingual [5] or in code-mixed language [6]. Monolingualism means the use of a single language, whereas code-mixed language means the use of two or more languages in speech or writing. Enough research is done on handling such hate content from Indian monolinguals, but little work is done for code-mixed languages and details about this is given in the next section.

A. Literature Study

The Table 1 shows the steps researchers used to create the code-mixed dataset. We found that almost all researchers have used social media platforms for data collection as they are the primary originators of code-mixed data. Most of them have thoroughly pre-processed the data because of its noisy nature. Some of the researchers evaluated the Inter-annotator agreement. We used Marathi for dataset creation. Though a Marathi-English dataset is available, it has a limited number of records with different sentiments. Marathi is the third-largest native speaker in India, the official language of Maharashtra state & has the richest and oldest literary tradition. The contribution of this paper is that we present the first gold standard code-mixed Marathi-English dataset having hate/offensive and not-hate/not-offensive sentiments. We also conducted an experimental analysis on our dataset for sentiment classification, employing various machine learning models.

II. Corpus Construction and Annotation

To collect a reasonable amount of Marathi-English comments, we used a YouTube platform. We have selected political incidents happened in Maharashtra in June & July of the year 2022 regarding "Eknath Shinde's Rebellion" to "Eknath Shinde's Speech in Legislative Assembly as Chief Minister". For collecting comments, we selected a team of 20 undergraduate students. A brief idea about data collection is given to these students and was provided with a Python script [17] for extracting the comments. The script extracts comments from each video and stores it into a Google sheet. Thus, a total of 21,606 comments from 95 YouTube videos were extracted. Later, these comments and replies of different Google sheets were merged into one excel file. We pre-processed the excel file by removing the personal information of the author who commented or replied to protect the author's privacy. Along with this, other information like Time, Likes, Reply Count, published, and updated were also removed, thus we got the "Comment" column. The duplicate comments, hashtags, hyperlinks and html tags were also removed. The HTML Character Entity references were converted to character references. The comments less than 5 words and greater than

TABLE I
DATASET AVAILABLE IN CODE-MIXED LANGUAGES

Dataset Name	**Social media platform**	**Pre-processing on collected data**	**Methodology for Inter-annotator agreement**	**class / labels Classes are balanced?**	**Dataset Size**
Hindi-English [7]	Facebook	Roman script comments, English sentences, comments having more than one sentence & comments over 50 words were eliminated.	02 persons Annotated. Cohen's Kappa coefficient	Positive, Negative, Neutral {Highly Imbalanced}	3879
Hindi-English & Bengali-English [8]	Twitter4j API	Incomplete, spam, tweet not having Bengali or Hindi words were removed. Hashtags & URLs are kept.	Manually annotated	Positive, Negative, Neutral	17921 & 5538
Bengali-English [9]	Twitter4j API	Tweet having minimum length as 8 & minimum 5 Bengali words were kept. Spam, incomplete & tweets with conflict sentiments were removed.	Manually annotated	Positive, Negative, Neutral	5000
Hindi-English [10]	Twitter	Removing URLs, Punctuations & replacing User Names and Emoticons	Manually annotated Cohen's Kappa coefficient	Hate speech, Normal speech {Classes are Imbalanced}	4575
Kannada-English [11]	Facebook	Comments in native script & mixed script were removed	Manually annotated	Positive, Negative, Neutral {Highly Imbalanced}	7005
Hindi-English [12]	Twitter	URLs, punctuations, user mentions & numbers were removed, hash tags & emoticons were converted into its text, text is lowercased & transliteration & translation is done.	Manually Annotated Cohen's Kappa coefficient	Non-offensive, Abusive, Hate-inducing {Severely Imbalanced }	3189
Malayalam-English & Tamil-English [13]	YouTube & Helo App	No details were mentioned.	Krippendorff's alpha	offensive, not-offensive {No Class Imbalance}	Both have size of 4000
Malayalam-English & Tamil-English [14]	YouTube	Comments other than code-mixed were totally removed, Comments with 5 to 15 words were selected, Emoticons, emoji's were removed	Manually Annotated Krippendorff's alpha	Positive, Negative, Neutral, Mixed feeling, Other language {Severely Imbalanced}	6739 & 15744
Telugu-English [15]	Twitter & Facebook	URLs, markup text & comments less than 5 words were removed	Manually Annotated Cohen's Kappa Coefficient	Positive, Negative, Neutral {Neutral class has less records}	19857
Marathi-English [16]	WhatsApp & YouTube	Special characters, emojis, & punctuations were removed	Manually Annotated	Positive, Negative, Neutral	1009

50 words were removed. Out of district emoticons, only one is kept, and the emoticon is converted into its text or meaning. Then we transliterated (Romanized) the comments using the indic_transliteration module. Indic_transliteration supports 09 Indian languages and 08 romanizations. The comments other than code-mixed were removed manually. Thus, after pre-processing, 4978 comments remained out of 21606.

A. Annotating the data into class

The annotation is done by a team of six people. Among them, three are undergraduate students and three are assistant professors of Computer Science. The first language of annotators is Marathi for speaking and writing and people who can also speak and write English. For annotating Tortus package is used. Students and professors are already familiar with the Python programming language as part of the curriculum. We provided the above Excel file, the necessary Python code and the training of annotation process for all annotators. They were asked not to be partial during the annotation process because comments are from a political background and their ideology can affect the annotation process. Annotators were informed about the annotation schema & have to annotate the comment into either an offensive/hate or not-offensive class. The annotation schema is as below.

- offensive/hate: Comment containing hate/offence/profanity aimed at an individual or group.
- not-offensive: Comment not containing any hate/offence/profanity.

After annotating the comments, Inter-annotator agreement (IAA) is calculated, which tells how well annotators can have agreement or disagreement on the dataset. We used Krippen-

TABLE II
DATA DISTRIBUTION

Label/Class	Marathi-English
offensive/hate	2450 (49.21%)
not-offensive	2528 (50.78%)
Total Comments	4978

TABLE III
CORPUS STATISTICS

Statistics	Values
Number of collected comments	21606
Annotated / Selected comments	4978
Number of special characters in the comments	2662
Number of words (without special characters)	59542
Number of words (with special characters)	62204
Unique words in dataset	15360
Average number of words per sentence (w/o special characters)	12
Average number of sentences per comment	1

dorff's alpha . It is a statistical measure of agreement among annotators to answer how much the resulting data can be relied upon to represent real data. Though it is computationally complex, in our case it is more relevant because we have used more than two annotators. The range of Krippendorff's alpha is between 0 and 1. Our annotation produced an agreement of 0.91.

B. Statistics and Facts about Corpus

Table 2 presents the label-wise distribution of the code-mixed Marathi-English dataset, in which offensive/hate and not-offensive sentiments were equally balanced. Table 3 shows the details of an enormous corpus with a vocabulary of 59452 words having 15360 unique words with an average of 12 words per comment. Table 4 shows some examples of the miscellaneous ways of writing in the dataset. We extracted the hate/offensive and non-hate words from the dataset and displayed them in Fig. 1

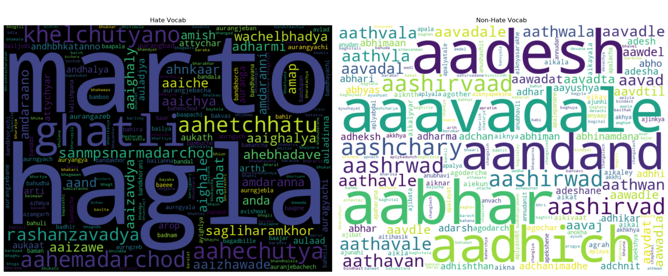

Fig. 1. Wordcloud of Hate and non-hate vocabulary

TABLE IV
MISCELLANEOUS WRITING IN DATASET

Dhanyawad, Dhanywaad, Dhanywad (Thanks)	Variations in Spelling
Zidabad, Zindabaad, Zindabad (Long Live)	
Vishvasaghata, Vishwasghaat (Betrayal)	
Nirlaj, Nirlajjya, Nirlajya (Shameless)	
Hijde, Hijade, Hijhade (Hijras)	
Adarchod, Madarachoda, Madarchiad (Motherfucker)	
Fadanvi, Fadanvisne, Fadnvisa, Farnadwis, Fadvinsh, …..	Variations in Nouns
Rout, Rauta	
Fadanvis-tarbooj, Sharad Pawar-wakdya, Chandrakant Patil-Champa	Politicians tease with different names

III. CHALLENGES WHILE ANNOTATING

The process of annotating the comments proved to be challenging due to disagreements among the annotators. The reason for this disagreement is that the comments come from a political background, and despite strict instructions to remain impartial, each annotator's ideology may have influenced their annotations. We provided some of the comments that exhibited disagreement.

- Marathi-English: ..quarter marun aalet cm saheb..
  English Meaning: CM has come drunk
  The above comment pertains to Eknath Shinde's speech in the Legislative Assembly. Despite the comment being explicitly offensive, one of the annotators perceived it as a comparison to the Chief Minister's speech resembling that of a drunk person.
- Marathi-English: tu la kahich mahit nahi mag.. obc cha hindutwa vegla ani, brahman cha hindutwa vegla
  English Meaning: You don't know anything then... OBC Hindutva is different and Brahmin Hindutva is different
  This comment means different groups and individuals may have various interpretations and perspectives on the concepts of OBC Hinduism and Brahmin Hinduism. Some annotators may come from diverse backgrounds and hold different beliefs hence they considered it's targeting to specific castes hence annotated as offensive.
- Marathi-English: hindi rastrabhasha nahi.. jara abhyas karun mahiti ghya.
  English Meaning: Hindi is not national language.. Just study and get information
  The above comment is answer to debate on whether Hindi should be considered as national language of India. Due to diverse viewpoints on this issue annotators have diverse opinion.

IV. BENCHMARK SYSTEM

On the newly developed dataset we applied different machine learning classifiers such as Support vector machine (SVC), Multinomial Naive Bayes (MNB), Logistic regression (LR), Random Forest (RF) and Decision Tree (DT). These classifiers were trained using CountVectorizer and TF-IDF

TABLE V
RESULT : MLC USING COUNTVECTORIZER

Classifiers	Class/ Sentiments	Word n-gram (1,5)		Character n-gram (1,5)		Combined Word & Character n-gram	
		F1-Score	Acc.	F1-Score	Acc.	F1-Score	Acc
Bag of Words - CountVectorizer							
SVC	**not-hate/ not-off.**	0.70	0.66	0.72	0.70	0.72	0.70
	hate/ offensive	0.61		0.69		0.68	
MNB	**not-hate/ not-off.**	0.70	0.68	0.74	0.73	0.75	**0.73**
	hate/ offensive	0.65		0.71		0.70	
LR	**not-hate/ not-off.**	0.69	0.67	0.73	0.72	0.73	0.71
	hate/ offensive	0.65		0.71		0.69	
RF	**not-hate/ not-off.**	0.71	0.65	0.71	0.68	0.72	0.69
	hate/ offensive	0.54		0.63		0.64	
DT	**not-hate/ not-off.**	0.64	0.61	0.61	0.60	0.61	0.59
	hate/ offensive	0.58		0.60		0.58	

TABLE VI
RESULT : MLC USING TF-IDF

Classifiers	Class/ Sentiments	Word n-gram (1,5)		Character n-gram (1,5)		Combined Word & Character n-gram	
		F1-Score	Acc.	F1-Score	Acc.	F1-Score	Acc
TF-IDF							
SVC	**not-hate/ not-off.**	0.70	0.67	0.75	**0.74**	0.74	0.72
	hate/ offensive	0.64		0.72		0.71	
MNB	**not-hate/ not-off.**	0.72	0.67	0.76	0.71	0.75	0.71
	hate/ offensive	0.62		0.63		0.64	
LR	**not-hate/ not-off.**	0.70	0.67	0.74	0.72	0.75	0.73
	hate/ offensive	0.63		0.70		0.71	
RF	**not-hate/ not-off.**	0.65	0.63	0.61	0.59	0.65	0.62
	hate/ offensive	0.62		0.57		0.59	
DT	**not-hate/ not-off.**	0.60	0.59	0.56	0.54	0.59	0.57
	hate/ offensive	0.58		0.52		0.55	

feature extraction techniques. We split the dataset into 70% 30% and evaluated results in terms of precision, recall and accuracy using sklearn. For both feature extraction techniques, we have used character n-grams, word n-grams and combined word and character n-grams. We tested the n-gram range from (1, 1) to (1, 5) and found that the classifiers were producing the best result for the n-gram range (1, 5) and only these results were mentioned. Table 5 shows the results of machine learning classifiers trained using CountVectorizer and Table 6 shows results of machine learning classifiers trained using TF-IDF. SVC produced the best accuracy of 74% with TF-IDF character n-gram features.

V. CONCLUSION

This paper introduces a code-mixed Marathi-English corpus, comprising YouTube comments that have been annotated for hate speech sentiment detection. The paper embraces an inter-annotator agreement score measured using Krippendorff's alpha and baseline results. The primary goal of this annotation project is to facilitate research on code-mixed sentiment detection and offer valuable and substantial amounts of data.

Additionally, the corpus is made openly available to the research community.

References

[1] https://timesofindia.indiatimes.com/city/mumbai/maharashtra-cops-file-over-600-cases-for-offensive-social-media-posts/articleshow/92992643.cms

[2] https://www.indiaspend.com/child-lifting-rumours-33-killed-in-69-mob-attacks-since-jan-2017-before-that-only-1-attack-in-2012-2012/

[3] VICE World News. https://www.vice.com/en/article/v7gq7j/controversial-facebook-post-on-prophetmuhammad-sparks-violent-riots

[4] https://timesofindia.indiatimes.com/readersblog/ezhil/violence-in-social-media-and-its-negative-impact-30702/

[5] Macrory, G., 2006. Bilingual language development: what do early years practitioners need to know?. Early years, 26(2), pp.159-169.

[6] Muysken, Pieter. Bilingual speech: A typology of code-mixing. Cambridge University Press, 2000.

[7] Ameya Prabhu, Aditya Joshi, Manish Shrivastava, and Vasudeva Varma. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. arXiv preprint arXiv:1611.00472, (2016).

[8] Braja Gopal Patra, Dipankar Das, and Amitava Das. Sentiment analysis of code-mixed indian languages: An overview of sail code-mixed shared task@ icon-2017. arXiv preprint arXiv:1803.06745, (2018).

[9] Soumil Mandal, Sainik Kumar Mahata, and Dipankar Das. Preparing bengali-english code-mixed corpus for sentiment analysis of Indian languages. arXiv preprint arXiv:1803.04000, (2018).

[10] Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. A dataset of hindi-english code-mixed social media text for hate speech detection. In Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media, pages 36–41, (2018).

[11] K Shalini, HB Barathi Ganesh, M Anand Kumar, and KP Soman. Sentiment analysis for codemixed Indian social media text with distributed representation. In 2018 International conference on advances in computing, communications and informatics (ICACCI), pages 1126–1131. IEEE, (2018).

[12] Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. Did you offend me? classification of offensive tweets in hinglish language. In Proceedings of the 2nd workshop on abusive language online (ALW2), pages 138–148, (2018).

[13] Bharathi Raja Chakravarthi, Anand Kumar M, John P McCrae, B Premjith, KP Soman, and Thomas Mandl. Overview of the track on hasoc-offensive language identification Dravidian codemix In FIRE (Working Notes), pages 112–120, (2020).

[14] Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. Language Resources and Evaluation, pages 1–42, (2022).

[15] Siva Subrahamanyam Varma Kusampudi, Preetham Sathineni, and Radhika Mamidi. Sentiment analysis in code-mixed telugu-english text with un- supervised data normalization. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pages 753–760, (2021).

[16] Patwardhan, Varad, et al. "Analysing The Sentiments Of Marathi-English Code-Mixed Social Media Data Using Machine Learning Techniques." 2023 International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE, 2023."Arts, South Asian". Encyclopædia Britannica. Encyclopædia Britannica 2007 Ultimate Reference Suite.

[17] https://github.com/MAN1986/LearningOrbis/blob/master/scrapeCommentsWithReplies.gs

9

Natural Language Processing in Healthcare: A Systematic Review

A systematic technique for reporting items was chosen by systematic reviews and meta-analyses to identify current clinical natural language processing (NLP) systems, which produce structured information from unstructured free text. The study collected information on the natural language processing frameworks, techniques, strategies, and practices used in healthcare applications. In order to find literature concerning NLP in healthcare, we used dependable indices such as Google Scholar, Scopus and Web of Sciences. We searched for journal articles and conference proceedings that were published between 2005 and 2020. Articles focusing on NLP in the healthcare system were selected from the available sources. Based on their emphasis on fruitful studies, 40 research articles were appraised. There were 19 publications on methodology, three on frameworks, five on approaches, five on processes and eight on review research papers. The NLP systems covered in this chapter could be used for a variety of clinical and research goals. In this study, we searched for NLP systems that have attempted to address issues such as "processing clinical free text and creating structured output". The information acquired from the highlighted research was examined to prioritize new approaches and challenges in clinical NLP.

9.1. Introduction

NLP is in high demand due to its undeniable potential for deciphering complicated, unstructured datasets and producing useful intelligence. Any type of information can be used, including text, speech and pictures. By combining this power, the organization can maximize its combined commitment in terms of

Chapter written by Bhanudas Suresh PANCHBHAI and Varsha Makarand PATHAK.

resources (cash, labor and time) and open doors to previously unattainable prospects. NLP facilitates the processing of enormous amounts of data supplied in a general linguistic format and the application of top-notch machine learning algorithms to it, in order to derive crucial business insights.

NLP is even more useful in the healthcare industry because there is a constant flow of enormous volumes of data there. Areas of healthcare that technology is altering include free-text, enhanced clinical documentation, data mining research, automated reporting, clinical trials and choices (Gupta et al. 2021). It has been claimed that improper use of medical terminology in the healthcare industry has created a number of problems with regards to good patient–provider communication (Dahm 2012). According to Keifenheim et al. (2015), there is a strong connection between the patient's question and their communication style. Healthcare personnel face very context-specific communication issues when speaking with patients, particularly when those patients converse in a language other than English (e.g. Marathi). When patients arrive at the emergency department (ED), their health-related complaints are noted. The hospital database contains the recorded data in an unstructured free-text format. In order to categorize and analyze the symptoms that patients have mentioned, healthcare practitioners retrieve this data (Wagholikar et al. 2011; Cyrus 2014). However, any inconsistency in the data that has been reported brings great medical shame. These circumstances are more frequent when the patient's language is unknown to the clinicians, according to Silverman et al. (2016). The signs and symptoms of an illness can be noted in a variety of ways. For instance, when a patient complains of chest pain, the symptoms can include tightness or discomfort in the chest, heartburn, pleurisy pain and angina. These signs and symptoms have a significant impact and serve as the main source of data for accurate diagnosis. When a human is required to perform these activities, grouping the patient's symptoms and diagnosing the disease based on these signs can present various difficulties. Meuter et al. (2015) suggest a method for describing a patient's symptoms that can be extremely subjective because it uses a lot of medical terminology. The diagnosis procedure can also become more difficult if common clinical terminologies are not used to describe the patient's concerns. If the patient's language is difficult for medical practitioners to understand, these complications could worsen. Manual translations of symptoms from the patient's native tongue to English are ineffective because there is a high likelihood of misunderstanding, and they are prone to errors.

The majority of the recorded clinical data exists as free text that is unstructured, making it challenging to understand. The following benefits can be attained by properly converting unstructured data to structured data:

1) This data is used in a secondary manner for extensive automated processing (Kreimeyer et al. 2017).

2) A reduction in the amount of time needed for manual expert review (Van Rosse et al. 2016).

Young et al. (2018) describe contemporary developments in NLP methodologies that have embraced a variety of cutting-edge techniques, such as machine learning and deep learning methods for transforming unstructured data. It is evident from the body of existing literature that deep learning techniques have surpassed machine learning techniques in terms of accuracy and processing capacity. According to Esteva et al. (2019), the majority of deep learning techniques are developed using supervised learning techniques. The corresponding models are developed to efficiently map the disorders and translate the raw data into pertinent symptoms using particular medical language: convolution neural networks (CNNs) and recurrent deep CNN. In NLP applications relevant to the healthcare industry, neural networks (RNN) are the most used methods. RNN models with a strong application potential in the healthcare industry include long short-term memory (LSTM) and reinforcement learning (RL).

The goal of this review of related literature on the use of NLP in healthcare is to examine current approaches, tools and methodologies. Along with identifying the NLP-in-healthcare method, the researchers also identified research issues pertaining to NLP applications in the field of medicine. The relevant medical diagnostic terminology is extracted from clinical notes and patient complaints using a configurable language processing system. Similarly, the underlining survey of relevant research is evaluated and offered in terms of recommendations and in terms of the outcome obtained with regards to the NLP strategy used.

9.2. Materials and methods

In order to address the following study questions, the researcher evaluated the literature:

1) What techniques, algorithms, devices, and tactics are currently being applied to integrate NLP into healthcare?

2) What NLP strategy was used to address the healthcare issue?

3) What makes the research proposal unique?

4) In terms of outcomes, what advice do we have regarding the NLP technique?

The review of this study was conducted according to the Kitchenham (2004) recommendations. In this situation, conducting a systematic review involves three main procedures: (1) develop a review plan, (2) execute the exam and (3) finish and submit the evaluation. Figure 9.1 shows Kitchenham's review methodology. At the planning stage – the review's identification – the research questions' specification and the review's methodology must be established. The goals of the study were predetermined before it began. We used Kitchenham's (2004) method for conducting this review. It consists of a set of excellent engineering principles for researchers that are universally accepted. The methodology of this review was based on this research strategy. This details the method used for the literature search, study selection and source selection. The review report technique also uses the framework outlined in Kitchenham's (2004) approach.

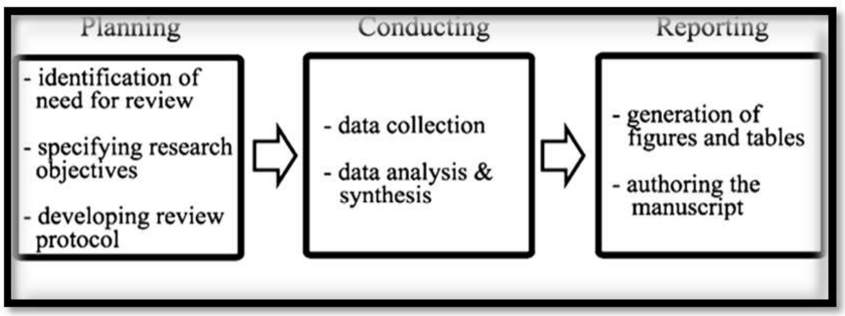

Figure 9.1. *Methodology for the review*

9.3. Data sources and searches strategy

For the period 2005–2020, conference proceedings and journal articles were searched in popular directories, including Web of Sciences, Google Scholar and Scopus. The phrase "Natural language processes in Health Care" was the first to be used as a string word for literature searches. The literature search also made use of abstracts, keywords and the title of the publication. In Table 9.1, for journal articles and the quantity of papers from a particular publication, and in Table 9.2 for conference proceedings, the sources of the literature search that was conducted using the aforementioned search string are found. The authentic papers to be included were chosen after a manual evaluation of all qualifying publications. A description of the methods and contributions of the various participants is provided in Table 9.3.

Serial no.	Journal name	No. of articles from the source
1	*JAMIA: Journal of the American Medical Informatics Association*	1
2	*Sage Journal: Health Informatics Journal*	1
3	*AMIA: Annual Symposium Proceedings Archive*	1

4	*National Library of Medicine PubMed*	20
5	*Research India publications: Advances in Computational Sciences & Technology*	1
6	*ITHEA@ Business and Engineering Applications of Intelligent and Information Systems*	1
7	*Journal of Biomedical Informatics*	5
8	*IEEE Computational Intelligence Magazine*	1
9	*International Journal of Computer Science and Information Security*	1
10	*Research India Publications: Journal of Theoretical and Applied Information Technology*	1
11	*Nature of Medicine*	1
12	*International Journal of Applied Engineering Research*	1
13	*International Journal of Computer Science Trends and Technology (IJCST)*	1
14	*International Journal of Medical Informatics*	1
15	*JMIR Medical Informatics*	1
16	*BMC Medical Informatics and Decision Making*	2
17	*Journal of Software Engineering & Applications*	1
18	*Journal of Hospital Librarianship*	1
19	*International Journal of Nursing Studies*	1
20	*International Journal of E-Health & Medical Communications*	1
21	*Journal of Theoretical and Applied Information Technology*	1

Table 9.1. *Literature sources from selected journals*

Serial no.	Conference proceedings name	No. of articles from the source
1	*Annual International Conference IEEE Engineering Medical Biological Science*	01
2	*Conference: Medical Informatics Europe (MIE)*	1
3	*International Conference on Analysis of Images, Social Networks & Texts*	1

4	*Conferences in Research and Practice in Information Technology*	1
5	*ACSW Frontiers 2007. The Australasian Workshop on Health Knowledge Management and Discovery*	1
6	*International Congress on Image & Signal Processing Biomedical Engineering & Informatics (2019)*	1
7	*8th ACM International Conference on Bioinformatics Computational Biology & Health Informatics (2017)*	1

Table 9.2. *Literature sources from conference proceedings*

9.3.1. *Requirements for inclusion*

The authors selected studies that examined the application of NLP to medical diagnosis and care. In order to determine whether the publications match the research questions outlined in the materials and methods for this study, they were examined and evaluated over the duration of 16 years.

9.3.2. *Exclusion standards*

During the literature search, the research papers were restricted, filtered and given boundaries using the following criteria. The range for the publication year was 2005–2020. Currently, only journal articles and conference papers written in English are taken into account. Only journal articles and conference papers were allowed to be included in the publication document type.

9.3.3. *Study selection*

We searched PubMed for any full-text English-language case reports, clinical trials and original research articles that used the phrase "Natural language processing in healthcare" to investigate how natural language processing (NLP) is being used in healthcare by researchers. We found 90 articles to use after removing duplicates from 125 citations that were pertinent to a string. After applying inclusion criteria, we finally discovered 40 research publications. It was vital to confirm the research papers' findings to make sure they provided the necessary answers to the questions raised during the 16-year investigation. Figure 9.2 shows a flowchart that details article extraction, screening and inclusion for meta-analyses and schematic reviews.

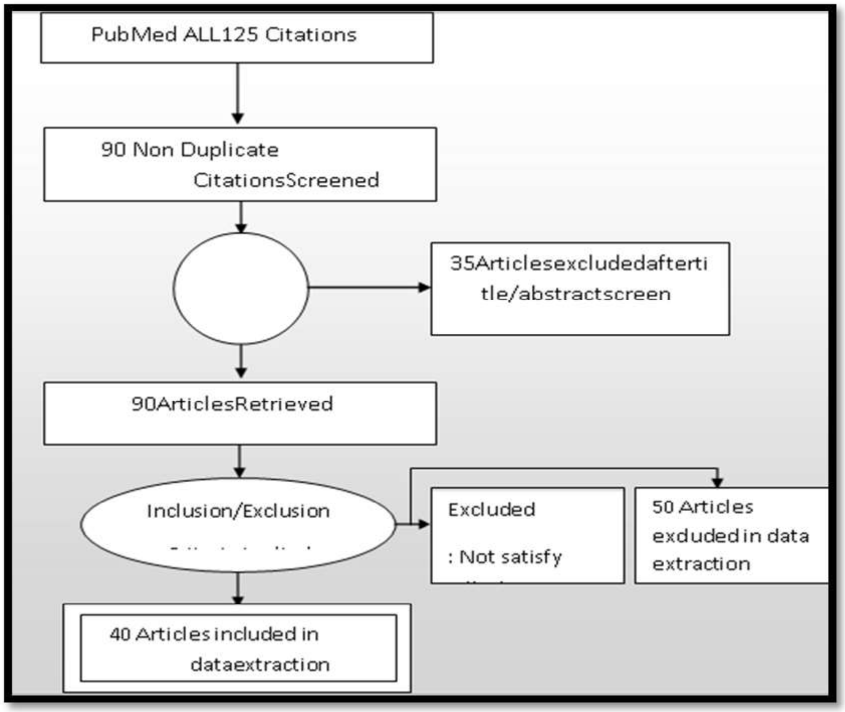

Figure 9.2. *Flowchart that details article extraction, screening and inclusion*

9.3.4. *Data extraction and synthesis*

The problem, methodology, data source and significant contributions that were used to establish NLP in healthcare are included in the information retrieved from the studies. Identification of research, inspiration, methodologies and outcomes were the types of information that were defined to carry out this study. The methodology is extracted with the ontology design approach as the foundation, as well as the research topic, applied techniques and their recommendations. The NLP approach in work and application focal point research is the foundation for the result consideration. The following are the main factors that we will discuss: system name and problem addressed data source, methodology and performance assessment for review papers on NLP in healthcare. Table 9.3 presents the summary of system parameters mentioned above and the contributions considered.

System and references	Author with year	System name	Problems attended	Source of data	Approach	Level of language status	Performance evaluation
System1	Chapman et al. (2005)	Classifying free-text triage chief complaints into syndrome categories with natural language processing	An application for classifying chief complaints into syndrome categories is presented	800 chief complaints	Used a natural language processing text classifier	International	Accuracy =90% Precision of 0.97 and 0.96
System2	Patrick et al. (2007)	SNOMED Clinical system	Automatic conversion of free text into a medical ontology	Real-time data Clinical notes	A medical concept from the SNOMED clinical terminology that can be identified automatically	International	Performance was within acceptable time and accuracy constraints
System3	Patrick et al. (2007)	SNOMED Clinical information management	To translate free text clinical notes into medical terminology and perform simple term composition	Electronic medical record of patients	Use the core algorithm token matcher for mapping text to the SNOMEDCT terminology	International	The system performed within acceptable time and accuracy constraints

System4	Dara et al. (2008)	Evaluation of preprocessing techniques for chief complaint classification	To determine whether preprocessing chief complaints automatically classifying them into syndrome categories improves classification performance	28,990 chief complaints	Use of two preprocessor: 1. Chief complaint processor 2. Emergency medical text processor	International	Accuracy=85%
System5	Apostolova et al. (2009)	Automatic segmentation number of clinical texts	This study attempts to automatically segment medical reports into semantic sections	A dataset of 215,000 free-text radiology reports	Use of the baseline algorithm and support vector classifier	International	Accuracy=90%

System6	Xu et al. (2010)	A medication information extraction (IE) system for clinical narratives	The present system extracts medication information from clinical notes	50 discharge summaries and clinic visit notes	A medication representation model	International	F-measure=93.2%
System7	Kaurova et al. 2011	Classification of free-text clinical narratives	The current study aims to present SLR of academic articles on clinical text classification in published from January 2013 to January 2018	Use of 72 primary studies from 8 bibliographical databases	Use of sampling methodologies, feature engineering, machine learning (ML) algorithms and performance measures	International	Review paper is good in the breadth and accuracy of the discussion
System8	Kreimeyer et al. (2017)	Natural language processing systems for capturing and standardizing unstructured	To perform systematic review on Natural Language Processing	Review of 71 different clinical natural language processing (NLP) systems	Query text NLP and structural data for inclusion and exclusion	International	Review paper is good in the accuracy of the discussion

		Clinical information: a systematic review	Systematic reviews and meta-analyses		Criteria		Clinical information A systematic review.
System9	Rahimi et al. (2012)	Developing ontology for data quality in chronic disease	Improving the data quality (DQ) of routinely gathered data for clinical care and research can help to enhance decision-making, evidence-based care and patient outcomes	Chronic disease data	Its ontology goes through five stages: specification, conceptualization, formalization, implementation and maintenance	International	In the electronic practice-based research network (ePBRN), build an ontological method for developing the3C of DQ for diabetes treatment

System10	Conway et al. (2013)	A review of chief complaint-based classifiers in North America	Information technology systems can use the automatic extraction of data from free text patient records to perform syndrome surveillance	15 paper reviews	Statistical approaches and keyword-based strategies are both used	International	This research examines 15 North American systems
System11	Mowafi et al. (2019)	A priority or global emergency care research in low-income countries	The absence of research on emergency chief complaints globally – especially in low-income countries	Patient chief complaints	To map free-text strings to standard medical terminology, machine-based techniques	International	Propose a study agenda for chief complaints in low-resource settings

System12	Ashish et al. (2014)	The pathology extraction pipeline for IE from pathology reports	To create a system for extracting information from pathology reports with the purpose of storing the information in a research data warehouse	Pathology reports	The method is based on ML algorithm (i.e. sequence mapping)	International	Extraction of several fields from pathology reports with excellent accuracy
System13	Ni et al. (2014)	Increasing the efficiency of patient identification for clinical trials in the emergency department (ED)	To develop an automated eligibility screening (ES) approach for clinical trials in an urban tertiary care pediatric ED	Between January 1, 2010 and August 31, 2012, we gathered eligibility criteria for 13 diseases	Use the effectiveness of NLP, IE and ML techniques on real-world clinical data and trials	International	Researchers demonstrated that NLP-, IE- and ML-based automated ES could successfully select patients for clinical trials by using the text of trial criteria and the content of EHRs

System14	Bill et al. (2014)	Automated extraction of family history information from clinical notes	This paper describes the development and evaluation of a NLP module based on the unstructured information management application (UIMA) for automated extraction of family history information with functionality for identifying statements, observations, and prediction (“indicator phrases”)	Sample of clinical notes	Use of UIMA (Unstructured Information Management Architecture) is a framework for an aging unstructured data	International	The family history NLP system achieved F-scores of 66.9, 92.4, 82.9, 57.3, 97.7 and 61.9
System15	McMurray et al. (2015)	Ontologically modeling of electronic health information exchange	Ontology was designed to measure and visualize regional interoperability	Data from are regional health system	Using Protégé 4, a knowledge-based framework and open-source web ontology language editor	International	Ontology was designed to measure and visualize regional interoperability

System16	Li et al. (2016)	A practical tool to implement hospital-based syndrome surveillance	The system discusses and examines the use of a symptom-clicking-module (SCM) as part of a hospital-based syndrome monitoring program	A total of 1,730,797 patient encounters were recorded by SCM	The clinicians used SCM to keep track of all of the patients who came in, and the data was automatically compiled and transmitted in daily batches. Using pre-defined criteria, the symptoms were grouped into seven targeted syndromes, and statistical techniques were used to detect temporal anomalies in the data series	International	Accuracy=92.1%

System17	Tripathi and Deshmukh (2017)	Building a database-driven reverse medical dictionary	The goal of the project was to create a fully complete reverse medical lexicon in order to improve the efficiency of health treatment consultations. Through an intelligent healthcare system, consumers can get rapid guidance on their health difficulties using a reverse medical vocabulary	User input is then the number of symptoms. Total inputs: 50	Use of three modules divided into three parts: preprocessing, filtering and ranking	Maharashtra	Accuracy=92%

System18	Jernite et al. (2013)	Predicting chief complaints at triage time in the emergency department	This study describes a method that helps achieve this objective by creating an extended ontology of chief complaints and automatically predicting a patient's top complaint based on their vitals and then users' description of their state upon arrival	A dataset of 97,000 triage notes	Use of linear support vector machine	International	Proposed a system for predicting a patient's main complaints based on a description of their current condition is good enough
System19	Usui et al. (2018)	Extraction and standardization of patient complaints from electronic medication histories for pharmacy covigilance	The goal of this study was to develop a system for collecting and standardizing patient complaints from computerized medication histories gathered data in Japanese community pharmacy in order to identify potential adverse drug event signals	A dataset of 5,000 patient complaints	Use of search rules on morphological analysis and speech	International	System performance was .66 regarding precision, .63 in recall and .65 for the F-measure

System20	Horng et al. (2019)	Modern ontology of emergency department presenting problem	Numerous attempts have been made to create a standardized “presenting problem” or “chief complaint” list to characterize the nature of an ED visit	A total of 180,424 patient visits were included in the study	Use of the hierarchical presenting problem ontology	International	Ontology successfully captured structured data for 95.9%.
System21	Barathi Ganesh et al. (2020)	Natural language understanding for medical texts	Natural language understanding is one of the essential tasks for building clinical text-based applications	Digital data in the form of clinical reports	Vector pace models are used, as well as sequential modeling jobs	National	Performance of 93.8% as the F1 score for i2b2 clinical corpus and achieves 97.29% as the F1 score for GENIA corpus
System22	Qing et al. (2019)	Neural network-based method for medical text classification	The approach creates sentence representations by combining two or more sentences. Dividing the document into segments and then combining them into a document representation.	Use of medical records	Use of neural networks	International	In this research, researcher suggests a novel hierarchical neural network method for medical text classification. Accuracy =Good

System23	Rousseau et al. (2019)	Data from emergency department	Comparing documentation on relevant clinical information in electronic health record (EHR). Providing note to computed tomography (CT) order requisition, prior to ordering of head CT for ED patients presenting with headache	Electronic health records	Performed between April 1, 2013 and September 30, 2014 at an adult quaternary academic hospital	International	Accuracy=90%
System24	Steinkamp et al. (2020)	Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes	ML and NLP have great potential to improve IE within electronic medical records for a wide variety of clinical search and summarization on tools	Use of 1,108 discharge summaries	For a clinically motivated symptom extraction ask, we present a task definition and detailed annotation requirements	International	The system is highly customizable to individual workflows and allows each user to choose which data should be structured and which should be unstructured

System25	Klyshinsky et al. (2020)	Formalization of medical records using ontology: patient complaints	Medical records contain a textual description of such important information as patients' complaints, diseases progress and therapy	Use of 100 medical records	To classify clinical statements into their assigned categories, a rule-based technique was applied	International	The algorithm corrects syntactical mistakes according to the hierarchical information from the ontology. The resulting algorithm was proved on 3,000 clinical records
System26	López-Úbeda et al. (2020)	An integrated approach to biomedical term identification systems	The authors present a unique architecture for developing biomedical term identification systems	Textual collections with clinical records	In order to construct the biomedical NER, we used certain NLP technologies. We begin by normalizing the content, which entails: removing punctuation, removing HTML elements, transforming the entire text to lowercase and coding it in UTF-8	International	BSB is an accurate system

System27	Yehia et al. (2019)	Ontology-based clinical IE from physician's free-text notes	An IE system that extracts structured data from handwritten clinical notes by physicians	The system is evaluated on real clinical notes	The OB-CIE system can help physicians to document visit notes without changing their workflow	International	F-measure of 94.90% and 97.80%
System28	Meystre et al. (2020)	Extracting information from textual documents in the electronic health record: a review of recent research	Examine recent published research on IE from textual documents in the electronic health record in this paper (EHR)	In this review, 174 papers were chosen and discussed	Literature review of the research published after 1995	International	Performance of IE systems with clinical text has improved since the last systematic review in 1995

System29	Lu et al. (2009)	Multilingual chief complaint classification for syndrome surveillance: An experiment with Chinese chief complaints	To ease data collection and analysis for automated syndrome surveillance, Chief Complaints must be grouped into established syndrome categories	Using statistical approaches, a set of 470 Chinese keywords was retrieved from around 1 million Chinese CC data	A novel Chinese CC classification system is proposed here, based on a Chinese–English translation module and an existing English CC classification on method	International	Accuracy=90%

System30	Wang et al. (2019)	Using natural language processing techniques to provide personalized educational materials for chronic disease patients in China: development and assessment of a knowledge-based health recommender system	The goal of this project was to create a health recommenders system in China that would deliver relevant teaching materials for chronic disease patients and assess its effectiveness	50 patients will be tested, and 100 educational documents will be distributed	Ontology and numerous NLP approaches were used to create a knowledge-based recommender system	International	A novel Chinese C classification system leveraging a Chinese–English translation module is better than others
System31	Lua et al. (2018)	Ontology-enhanced automatic chief complaint classification for syndromic surveillance	A new ontology-enhanced automatic CC classification approach is presented in this paper. In a medical ontology, using semantic relations	Real-world dataset	The UMLS-based weighted semantic similarity score (WSSS) grouping mechanism was used	International	Our ontology-enhanced strategy out performs the benchmark methods in terms of sensitivity, F-measure and F2 measure, according to his study

System32	Hier and Brint (2020)	A neuro-ontology for the neurological examination	Based on UMLS Met Thesaurus concepts, we investigated the feasibility of recording the neurological examination as machine-readable codes	A dataset of 2,386 test-cases was constructed based on 419 published neurological illnesses	Using 1,100 concepts from the UMLS Met thesaurus	International	The neurology examination ontology (NEO), which was created by combining different terminologies in UMLS
System33	Mall and Jaiswal (2018)	Survey of machine translation for Indian languages to English and its approaches	The purpose of this paper is to discuss the various methodologies used in translation systems for Indian languages to English languages	A total of 16 research papers on the conversion of Indian languages to English have been published	List the numerous MT approaches for converting Indian languages into other languages, as well as their advantages and disadvantages	National	Accuracy=80% For Indian language Hindi. Chunk performance is improved

System34	Dara et al. (2008)	Evaluation of preprocessing techniques for chief complaints	To see whether preparing a chief complaint before automatically categorizing them into syndrome groups improves classification accuracy	Using two preprocessors, chief complaints were preprocessed (CCP and EMT-P)	We preprocessed chief complaints using two preprocessors (CCP and EMT-P) and evaluated whether classification performance increased for a probabilistic classifier (CoCo)	International	CCP exhibited High accuracy=85% (chief complaint preprocessing)

System35	Mall and Jaiswal (2018)	Using chief complaints for syndrome surveillance: a review of chief complaint-based classifiers in North America (Mall and Jaiswal 2018)	This article examines 15 syndrome monitoring systems in North America, including those in cities, countries, states and the federal government	Reviewed 15 research papers	The studies on classifiers can be classified into two categories: statistical methods and keyword-based methods	International	All of the systems examined can be linked to respiratory and gastrointestinal disorders
System36	Dara et al. (2008)	Clinical information extraction applications: a literature review	A survey of the literature for clinical data extraction applications	There are 263 publications that have been thoroughly reviewed	A literature search was conducted using Ovid MEDLINE In-Process and other non-indexed citations, Ovid MEDLINE, OvidEMBASE, Scopus, Web of Science, and ACM Digital Library for publications published between January 2009 and September 2016	International	For title and abstract screening, a total of 1,917 publications were identified. 263 articles were chosen and discussed in this evaluation from among these publications

System37	Hansoti et al. (2021)	Calibrating a chief complaint list for lower source settings: a methodological case study	This research was done as part of a wider prospective observational study on human immune deficiency virus testing in South African emergency departments	Paper case report forms were used to collect data on 3,357 patients	The frequency of concordance between the coded chief complaint word and the selected symptom(s) from the pilot symptom list was determined by two members of the study team	International	A systematic process for calibrating a chief complaint list or the local context was described in this study
System38	Musa et al. (2014)	Ontology knowledge map for enhancing healthcare services: a case of emergency unit of specialist hospital	It offers an ontology knowledge map-based strategy for locating superfluous us transactions that need to be modified in order to improve healthcare administration	Use of electronic health records	We chose the ontology as the study's base because it is thought to be capable of providing a better understanding of an organization dynamics, allowing for a good alignment between enterprise design and operation, and allowing for a systematic reengineer plan	National	One of the ways commended is to automate the emergency departments by introducing EHR systems, which will make it easier for the actors in the unit to rapidly and accurately obtain all of the information they need about a patient

System39	Elmessiry et al. (2017)	Medical diagnosis by complaints of patients and ML	The objective is to identify a link between the patient's complaints and probable diseases	Dataset of 10,000 authorial medical website	The key to detecting correlations between patients' complaints and probable diseases is to use ML models, as describe d in this research	International	Accuracy=75% Precision=81% Recall=81%
System40	Elmessiry et al. (2017)	Leveraging sentiment analysis for classifying patient complaints	The goal of this study is to automate the classification of patient complaints in order to enhance triage and response times	Use of electronic health records	Using increased linguistic inquiry and a word count lexicon, map each complaint to a vector	International	Accuracy=84%

Table 9.3. *Summary of approaches and contribution*

9.4. Results and discussion

19 of the 40 potential research projects were methodology-based, five were technique-based, three were framework-based, five were process-based and eight were assessments of existing projects. The methods taken into account are ontology-based, as shown in Table 9.3, and they involve the usage of the Protégé-owl editor tool, Protégé 4, OWL 2, OWL and SNOMED CT. The main contributions include, but are not limited to: modeling of protégé-based knowledge representation for linking concepts and data for diabetes diseases; mobile-based healthcare ontology; classification of diseases based on phenotypes; advancement in service provision and accessibility to trustworthy health data. Table 9.4 displays the findings of the review. A graphical representation is shown in Figure 9.3. This provides an overview of the number of studies per initiative that were located and shows the data about the functions intended to enhance healthcare operational procedures. These characteristics can serve as the foundation for approaches, frameworks, workflows, methodologies and reviews.

System number	System name	Types of initiative	Type of method/tech./process/frame/ review
System1	Syndrome Biosurveillance system (Chapman et al. 2005)	Method	Experimental
System2	SNOMED clinical system (Patrick et al. 2007)	Method	Experimental
System3	SNOMED clinical information management (Patrick et al. 2007)	Method	Experimental

System4	Evaluation of preprocessing techniques for chief complaint classification (Dara et al. 2008)	Process	Process mapping: use of two preprocessors (CCP and EMT-P)
System5	Automatic segmentation of clinical texts (Apostolova et al. 2009)	Method	Experiment setup: rule-based algorithm
System6	A medication information extraction system for clinical narratives (Xu et al. 2010)	Method	Mixed method
System7	Classification of free-text clinical narratives (Kaurova et al. 2011)	Review	Systematic review
System8	Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review (Kreimeyer et al. 2017)	Review	Systematic review
System9	Developing ontology for data quality in chronic disease (Rahimi et al. 2012)	Technique	Developed an ontological toolkit to support research and quality improvement studies

System10	A review of chief complaint-based classifiers in North America (Conway et al. 2013)	Review	Literature review
System11	A priority for global emergency care research in low-income countries (Mowafi et al. 2019)	Review	Systematic review
System12	The pathology extraction pipeline for information extraction from pathology reports (Ashish et al. 2014)	Process	Sequence mapping
System13	Increasing the efficiency of patient identification for clinical trials in the emergency department (Ni et al. 2014)	Technique	Use of leveraging natural language processing, information extraction and machine learning technologies
System14	Automated extraction of family history information from clinical notes (Bill et al. 2014)	Framework	Unstructured information management architecture
System15	Ontological modeling of electronic health information exchange (McMurray et al. 2015)	Framework	Conceptual framework

System16	A practical tool to implement hospital based syndrome surveillance (Li et al. 2016)	Process	Sequential mapping
System17	Building a database-driven reverse medical dictionary (Tripathi and Deshmukh 2017)	Method	Experimental
System18	Predicting chief complaints at triage time in the emergency department (Jernite et al. 2013)	Method	Experimental
System19	Extraction and standardization of patient complaints from electronic medication histories for pharmacy covigilance (Usui et al. 2018)	Method	Experimental
System20	Modern ontology of emergency department presenting problems (Horng et al. 2019)	Method	Classification
System21	Natural language understand for medical texts (Barathi Ganesh et al. 2020)	Framework	Linear

System22	Neural network-based method for medical text classification (Qing et al. 2019)	Method	Experimental
System23	Automated retrieval of data from emergency department (Rousseau et al. 2019)	Method	Observational
System24	Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes (Steinkamp et al. 2020)	Techniques	Develop model for symptom extraction from unstructured clinical notes
System25	Formalization of medical records using ontology: Patient complaints (Klyshinsky et al. 2020)	Method	Experimental
System26	An integrated approach to biomedical term identification systems (López-Úbeda et al. 2020)	Framework	Modular-based
System27	Ontology-based clinical information extraction from physician's free-text notes (Yehia et al. 2019)	Process	Rule-based

System28	Extracting information from textual documents in the electronic health record: A review of recent research (Meystre et al. 2020)	Review	Systematic review
System29	Multilingual chief complaint classification for syndromic surveillance: An experiment with Chinese chief complaints (Lu et al. 2009)	Method	Experimental
System30	Using natural language processing techniques to provide personalized educational materials for chronic disease patients in China: Development and assessment of a knowledge-based health recommender system (Wang et al. 2019)	Technique	Rule-based approach
System31	Ontology-enhanced automatic chief complaint classification for syndromic surveillance (Lua et al. 2018)	Method	This paper uses two popular CC classification methods using a real-world dataset
System32	A neuro-ontology for the neurological examination (Hier and Brint 2020)	Method	Use of ontology

System33	Survey of machine translation for Indian languages to English and its approaches (Mall and Jaiswal 2018)	Review	Review on 16 research papers
System34	Evaluation of preprocessing techniques for chief complaints (Dara et al. 2008)	Process	We preprocessed chief complaints using two preprocessors (CCP and EMT-P)
System35	Using chief complaints for syndromic surveillance: A review of chief complaint-based classifiers in North America (Mall and Jaiswal 2018)	Review	Review on 15 research papers
System36	Clinical information extraction applications: A literature review (Dara et al. 2008)	Review	Review 263 articles selected for review
System37	Calibrating a chief complaint list for lower source settings: A methodological case study (Hansoti et al. 2021)	Method	Paper presents methodological strategy that can be exported to other settings to refine a local chief complaint list
System38	Ontology knowledge map for enhancing healthcare services: A case of emergency unit of specialist hospital (Musa et al. 2014)	Method	Method-based ontology

System39	Medical diagnosis by complaints of patients and machine learning (Elmessiry et al. 2017)	Technique	Machine learning
System40	Leveraging sentiment analysis for classifying patient complaints	Method	Vector-based

Table 9.4. *Summary of quantity of studies per initiative*

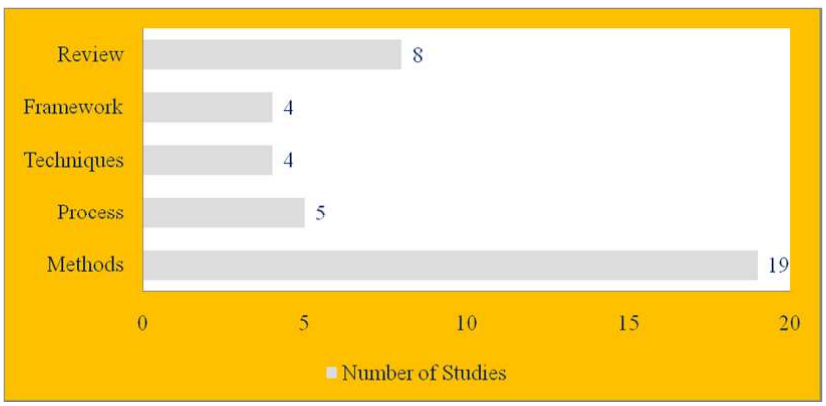

Figure 9.3. *Summary of quantity of studies per initiative discovered. For a color version of this figure, see www.iste.co.uk/pradhan/cognitivecomputing.zip*

9.5. Conclusion

The available literature on the use of NLP in healthcare was compiled in this review. This helps to provide researchers in the sector with complete information about currently used approaches, improvements to the caliber of healthcare services, and limitations. It included information on the goals of particular researchers as well as their contributions to enhancing healthcare through the use of NLP, and the restrictions placed on their research. It was also shown that several researchers in this field used a range of strategies to enhance healthcare through NLP. Some approaches used particular methodologies, techniques, procedures and frameworks, while others were modifications of those that previously existed. This healthcare

NLP review identified the requirement for ontology-based models to enhance the delivery of healthcare services for both patients and healthcare providers. Researchers in the field of NLP in healthcare can use this study's findings to pinpoint any gaps or problem areas. As a result, the research field will contribute even more.

9.6. References

Apostolova, E., Channin, D., Fushman, D.D., Lytinen, S.L., Stan Raicu, D. (2009). Automatic segmentation of clinical texts. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 5905–8. doi: 10.1109/IEMBS.2009.5334831.

Ashish, N., Dahm, L.M., Boicey, C.M. (2014). Pathology extraction pipeline: The pathology extraction pipeline for information extraction from pathology reports. *Health Informatics Journal*, 20(4), 288–305. doi: 10.1177/1460458213494032.

Barathi Ganesh, H.B., Reshma, U., Soman, K.P., Anand Kumar, M. (2020). Med NLU: Natural language understander for medical texts. In *Deep Learning Techniques for Biomedical and Health Informatics*, Dash, S., Acharya, B., Mittal, M., Abraham, A., Kelemen, A. (eds). Springer, Cham. doi: 10.1007/978-3-030-33966-1_1.

Bill, R., Pakhomov, S., Chen, E.S., Winden, T.J., Carter, E.W., Melton, G.B. (2014). Automated extraction of family history information from clinical notes. *AMIA Annu. Symp. Proc.*, 1709–1717.

Chapman, W.W., Christensen, L.M., Wagner, M.M., Haug, P.J., Ivanov, O., Dowling, J.N., Olszewski, R.T. (2005). Classifying free-text triage chief complaints into syndrome categories with natural language processing. *Artif. Intell. Med.*, 33(1), 31–40.

Conway, M., Dowling, J.N., Chapman, W.W. (2013). Using chief complaints for syndrome surveillance: Are view of chief complaint based classifiers in North America. *J. Biomed. Inform.*, 46(4), 734–743.

Cyrus, J.W. (2014). A review of recent research on internet access, use, and on line health information seeking. *Journal of Hospital Librarianship*, 14(2), 149–157.

Dahm, M.R. (2012). Tales of time, terms, and patient information-seeking behavior—an exploratory qualitative study. *Health Communication*, 27(7), 682–689.

Dara, J., Dowling, J.N., Travers, D., Cooper, G.F., Chapman, W.W. (2008). Evaluation of preprocessing techniques for chief complaint classification. *Journal of Biomedical Informatics*, 41, 613–623.

Elmessiry, A., Zhang, Z., Cooper, W.O., Catron, T.F., Karrass, J., Singh, M.P. (2017). Leveraging sentiment analysis for classifying patient complaints. *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 44–51.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., De Pristo, M., Chou, K., Dean, J. (2019). A guide to deeplearning in healthcare. *Nature Medicine*, 25(1), 24–29.

Gupta, A., Lai, A., Mozersky, J., Ma, X., Walsh, H., DuBois, J.M. (2021). Enabling qualitative research data sharing using a natural language processing pipeline for deidentification: Moving beyond HIPAA Safe Harbor identifiers. *JAMIA Open*, 4(3). doi: 10.1093/jamiaopen/ooab069.

Hansoti, B., Hahn, E., Rao, A., Harris, J., Jenson, A., Markadakis, N., Moonat, S., Osula, V., Pousson, A. (2021). Calibrating a chief complaint list for low resource settings: A methodological case study. *Int. J Emerg. Med*., 14(1), 32.

Hier, D.B. and Brint, S.U. (2020). A neuro-ontology for the neurological examination. *BMC Medical Informatics and Decision Making*, 20, 47.

Horng, S., Greenbaum, N.R., Nathanson, L.A., McClay, J.C., Goss, F.R., Nielson, J.A. (2019). Consensus development of a modern ontology of emergency department presenting problems. *Appl. Clin. Inform*., 10(3), 409–420.

Jernite, Y., Halpern, Y., Horng, S., Sontag, D. (2013). Predicting chief complaints at triage time in the emergency department. *NIPS 2013 Workshop on Machine Learning for Clinical Data Analysis and Healthcare*.

Kaurova, O., Alexandrov, M., Blanco, X. (2011). Classification of free text clinical narratives (short review). *Business and Engineering Applications of Intelligent and Information Systems*, 124.

Keifenheim, K.E., Teufel, M., Ip, J., Speiser, N., Leehr, E.J., Zipfel, S., Herrmann-Werner, A. (2015). Teaching history taking to medical students: A systematic review. *BMC Medical Education*, 15(1), 159.

Kitchenham, B. (2004). Procedures for performing systematic reviews. Technical Report, Keele University, 33.

Klyshinsky, E.S., Gribova, V.V., Shakhgeldyan, C., Shalfeeva, E.A., Okun, D., Geltser, B.I., Gorbach, T.A., Karpik, O.D. (2020). Formalization of medical records using an ontology: Patient complaints. In *Analysis of Images, Social Networks and Texts*, van der Aalst, W.M.P., Batagelj, V., Ignatov, D.I. et al. (eds). Springer, Cham, 143–153.

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S.F., Forshee, R., Walderhaug, M., Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics*, 73, 14–29.

Li, G., Song, H., Liang, H.-N., Qu, Y., Liu, L., Bai, X. (2019). Medical diagnosis by complaints of patients and machine learning. *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Suzhou, 1–5. doi: 10.1109/CISP-BMEI48845.2019.8965949.

López-Úbeda, P., Díaz-Galiano, M.C., Montejo-Ráez, A., Martín-Valdivia, M.T. (2020). An integrated approach to biomedical term identification systems. *Appl. Sci*., 10(5), 1726.

Lu, H.-M., Chen, H., Zeng, D., King, C.-C., Shih, F.-Y., Wu, T.-S., Hsiao, J.-Y. (2009). Multilingual chief complaint classification for syndrome surveillance: An experiment with Chinese chief complaints. *Int. J Med. Inform*., 78(5), 308–320.

Lua, H.-M., Zengab, D., Trujilloc, L., Komatsuc, K., Chena, H. (2008). Ontology-enhanced automatic chief complaint classification or syndromic surveillance. *J. Biomed. Inform*., 41(2), 340–356.

Mall, S. and Jaiswal, U.C. (2018). Survey: Machine translation for Indian language. *International Journal of Applied Engineering Research*, 13(1), 202-209.

McMurray, J., Zhu, L., McKillop, I., Chen, H. (2015). Ontological modeling of electronic health information exchange. *Journal of Biomedical Informatics*, 56, 169–178.

Meuter, R.F., Gallois, C., Segalowitz, N.S., Ryder, A.G., Hocking, J. (2015). Overcoming language barriers in healthcare: A protocol for investigating a few and effective communication when patients or clinicians use a second language. *BMC Health Services Research*, 15(1), 371.

Meystre, S.M., Savova, G., Kipper-Schuler, K.C., Hurdle, J.F. (2020). Extracting information from textual documents in the electronic health record: A review of recent research. *Yearb. Med. Inform*., 128–144.

Mowafi, H., Workis, D.D., Bisanzo, M., Hansoti, B., Seidenberg, P., Obermeyer, Z., Hauswald, M., Reynolds, T.A. (2019). Making recording and analysis of chief complaint a priority for global emergency care research in low-income countries. *Acad. Emerg. Med*., 20(12), 1241–1245.

Musa, M., Othman, M., Al-Rahimi, W.M. (2014). Ontology knowledge map for enhancing health care services: A case of emergency unit of specialist hospital. *Journal of Theoretical and Applied Information Technology*, 70, 196–209.

Ni, Y., Kennebeck, S., Dexheimer, J.W., McAneney, C.M., Tang, H., Lingren, T., Li, Q., Zhai, H., Solti, I. (2014). Automated clinical trial eligibility prescreening: Increasing the efficiency of patient identification for clinical trials in the emergency department. *J. Am. Med. Inform. Assoc*., 22(1), 166–78.

Patrick, J., Wang, Y., Budd, P. (2007). An automated system for conversion of clinical notes into SNOMED clinical terminology. *Proceedings of the fifth Australasian symposium on ACSW frontiers*, Ballarat, 68, 219–226.

Qing, L., Linhong, W., Xuehai, D. (2019). A novel neural network-based method for medical text classification. *Future Internet*, 11, 12.

Rahimi, A., Liaw, S.-T., Taggart, J., Ray, P., Yu, H. (2012). Developing ontology for data quality in chronic disease management. *International Journal of E-Health and Medical Communications*, 5(3), 58–77. doi: 10.4018/ijehmc.2014070105.

Rousseau, J.F., Ip, I.K., Raja, A.S., Valtchinov, V.I., Cochon, L., Schuur, J.D., Khorasani, R. (2019). Can automated retrieval of data from emergency department physician notes enhance the imaging order entry process? *Appl. Clin. Inform*., 10(2), 189–198.

Silverman, J., Kurtz, S., Draper, J. (2016). *Skills for Communicating with Patients*. CRC Press, Boca Raton.

Steinkamp, J.M., Bala, W., Sharma, A., Kantrowitz, J.J. (2020). Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes. *J. Biomed. Inform*., 102, 103354. doi: 10.1016/j.jbi.2019.103354.

Tripathi, P. and Deshmukh, M. (2017). Building a scalable database driven reverse medical dictionary using NLP techniques. *Advances in Computational Sciences and Technology*, 10(5), 1221–1231.

Usui, M., Aramaki, E., Iwao, T., Wakamiya, S., Sakamoto, T., Mochizuki, M. (2018). Extraction and standardization of patient complaints from electronic medication histories for pharma covigilance: Natural language processing analysis in Japanese. *JMIR. Med. Inform*., 6(3), e11021.

Van Rosse, F., de Bruijne, M., Suurmond, J., Essink-Bot, M.L., Wagner, C. (2016). Language barriers and patient safety risks in hospital care. A mixed methods study. *International Journal of Nursing Studies*, 54, 45–53.

Wagholikar, A.S., Lawley, M.J., Hansen. P., Chu, K. (2011). Identifying symptom groups from emergency department presenting complaint free text using SNOMED CT. *AMIA Annual Symposium Proceedings*, 1446.

Wang, Z., Huang, H., Cui, L., Chen, J., An, J., Duan, H., Ge, H., Deng, N. (2019). Using natural language processing techniques to provide personalized educational materials for chronic disease patients in china: development and assessment of a knowledge-based health recommender system. *JMIR. Med. Inform*., 8(4), e17642. doi: 10.2196/17642.

Xu, H., Stenner, S.P., Doan, S., Johnson, K.B., Waitman, L.R., Denny, J.C. (2010). MedEx: A medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc*., 17(1), 19–24.

Ye, C., Li, Z., Fu, Y., Lan, Y., Zhu, W., Zhou, D., Zhang, H., Lai, H.Z.S., Buckeridge, D.L., Sun, Q. et al. (2016). SCM: A practical tool to implement hospital based syndrome surveillance. *BMC Res Notes* 9, 315. doi: 10.1186/s13104-016-2098-z.

Yehia, E., Boshnak, H., Abdel, S., Amany, G., Doaa, A., Elzanfaly, S. (2019). Ontology-based clinical information extraction from physician's free-text notes. *Journal of Biomedical Informatics*, 98, 103276.

Young, T., Hazarika, D., Poria, S., Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3), 55–75.

IEEE Computational Intelligence Society | IEEE

HYDERABAD SECTION

IEEE Hyderabad Section

Intelligent Systems and Machine Learning Conference 2024 (ISML)

May 4-5th, 2024

IEEE CIS/GRSS Joint chapter of IEEE Hyderabad section

CERTIFICATE OF PARTICIPATION

CONGRATULATIONS & GRATITUDE TO

Rupali Vasant Narkhede

FOR PRESENTING PAPER

Marathi Language Query Interface (MLQI) for GIS database using OpenAI

MOUSMI AJAY CHAURASIA

CONFERENCE CHAIR

CHRIST(Deemed to be University) Pune, Lavasa Campus
&
Indian Institute of Information Technology Allahabad, Prayagraj

present this

CERTIFICATE OF PARTICIPATION

to

Jayashree Dhananjay Chaudhuri

For presenting a paper titled

Investor Personalities - Navigating Behavioral Biases Through the Lens of Traits

at 7th SSBF and International Conference on Behavioural Finance and other Contemporary Issues (ICBF-24) held on 17th February 2024, organized by the School of Business and Management, CHRIST(Deemed to be University) Pune, Lavasa Campus in Collaboration with Department of Management Studies, Indian Institute of Information Technology Allahabad, Prayagraj, India.

Prof Ranjit Singh
General Chair, ICBF' 24

Dr Fr Lijo Thomas
Dean & Director