**Different roles of Artificial Intelligence in Natural Language Processing**

|  |  |
| --- | --- |
| Ms. Pooja H. Rane | Dr. Pranjal S. Bogawar |
| Computer Science | Computer Science |
| Aakar College of Management for Women, Hingna, Nagpur, India | Aakar College of Management for Women, Hingna, Nagpur, India |
| rane.pooja20@gmail.com | pbogawar@gmail.com |

**ABSTRACT**

Artificial Intelligence is a branch of Science which can able to learn, make decisions, and take action even when it encounters a condition it has never come over before. In 1956, firstly Artificial Intelligence concept was introduced in computing conference by John McCarthy of Dartmouth University. In today’s world artificial intelligence is used everywhere like manufacturing, Health, cyber security, online shopping & advertising etc. Legal Artificial Intelligence is mainly focuses on applying artificial intelligence technology to help legal tasks. Natural Language processing is a branch of AI which works on analysis/ recognition of text in soft format, hard format, and audio format. The different resources majority in this field are presented in text forms, such as judgment documents, contracts, and legal opinions. Therefore, most Legal AI tasks are based on Natural Language Processing (NLP) technologies. In this survey paper we study different use of Natural Language Processing in Text Mining, Email-Mining; Speech Recognition etc. Artificial Intelligence is implemented using various algorithms like K-Means Algorithm, Support Vector Machine (SVM) algorithm, Machine Learning Algorithm, Hill- Climbing Algorithm, Supervised Learning Algorithm, Baum Welch Algorithm etc. The brief introduction of these algorithms is also provided in this paper.

**Keywords:** Text Mining, Email Mining, Phishing, Speech Recognition, Natural Language Processing,

**I INTRODUCTION**

Natural Language Processing **(**NLP) is a branch of computer science or artificial intelligence which works on different aspects like speech recognition, image processing, email mining, context free grammar, dependency grammar, data extraction, language translation, Multi page document classification ,resume, letters, official documents and voice data from different medium like social media, audio and video etc. Natural language interpretation or intangibility improves computers or machines or devices ability by learning information online and apply what they learned in the real world. This ability is useful for generating creative data, facts in very less amount of time. Natural language processing began in the 1940s. At this time, mostly people were working on translation of one language to another with the help of machine that could perform translation automatically [9].

In 1958, John McCarthy was identifying significant issues in the development of NLP. These issues are Language differences, Training Data, Phrasing ambiguities, Misspelling, words with multiple meaning etc. [10]. Around 1957-1970, researchers split into two divisions concerning NLP: symbolic and stochastic [10]. Symbolic, or rule-based, researchers focused on formal languages and generating syntax; this group consisted of many linguists and computer scientists who considered this branch the beginning of artificial intelligence research [10]. The symbolic expressions distributed expression in metric spaces where comparability includes in examples is used to learn frequency or particular tasks by using neural networks or other machine learning models. For example, two sentences such as s1 = “a mouse eats some cheese” and s2 = “a cat swallows a mouse” can be considered in many different ways: (1) Number of words in common; (2) realization of the pattern “ANIMAL EATS FOOD.”

Stochastic researchers were more interested in statistical and probabilistic methods of NLP, working on problems of optical character recognition and pattern recognition between texts [10]. In stochastic various machine learning models were used. These are Linear Regression, Logistic Regression, Support Vector Machines (SVMs), Decision Trees, and Random Forests (RFs) were implemented to tackle problem More recently, deep learning models including Artificial Neural Networks (ANNs) [1,2,3,11], Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs) [11].

In this paper we review various papers in Natural Language Processing and did the comparative studies.

**II LITERATURE REVIEW**

The natural Language Processing is used to analyze/recognize the written/image/audio material. Here we explore some of the work of NLP in various areas.

1. **Text Mining**

Text mining also called as data text mining which is transforming unstructured text into structured text format by identifying meaning pattern [2]. In 1989 Marti Hearst firstly used text data mining [4]. In text mining text is one most common data type in databases. According database data are categorized into two different forms i.e. structured data and unstructured data.

Structured data is use in the tabular format numerous in rows & column to process data easily process for machine learning algorithm. It includes different inputs like names, addresses, and phone numbers.

Unstructured data have a predefined data format. It can include text from sources, like social media or product reviews, or rich media formats like, video and audio files. In text mining various different algorithms were used such as extraction algorithm, K- Means algorithm, Support Vector Machine (SVM), Clustering algorithms, Hierarchical Clustering Algorithms. The text mining is mainly used to mine the information from

Tagging parts of speech, parsing syntax, tokenizing. The text mining gives suggestion for various words, their spelling, sentences, grammar in the sentences as well as utilized for latest topic plagiarism.

1. **Email Mining**

The email mining is different from text mining as it includes various fields such as: spam detection, Email categorization, contact analysis which is used for analysis of email. An email stands for electronic mail which is used to communicate through electronic devices to deliver message through computer network. Email has introduced in the year 1970 by Ray Tomlinson created a way to transmit messages between computer systems on the Advanced Research Projects Agency Network (ARPANET) [1]. The email is one of the most popular methods of digital communication. Now days emails are used for phishing [1-3], spam[4], domain spoofing [3], Sentiment analysis of the users[2 ], business email compromise (BEC)[3].

In the year 1990s first time someone used the term 'phishing' can be traced back and January 2nd, 1996. Hackers would affect to be America Online (AOL) administrators and phish for login credentials so they can access the internet for free. Phishing extracts sensitized to information from fraudulently attempting and is a social engineering threat [1]. Hence email mining is required to give idea to the users regarding these mails.  According to different proposed comparison phishing email detection uses NLP techniques. Generally phishing detection research has concentrating on procedure for automated phishing detection. Researches used hundreds of features for detection of phishing emails, spam emails. Here we are mentioning some of the important features used for identification of phishing/spam emails. These are as follows:

1. Email body-based characteristics: These qualities are taken from the e-mail body. They have double highlights like shapes, HTML, and particular expressions and joins within the mail body [1].

2. Subject-based features: Certain Perspective of an e-mail are determines from its subject, such as whether it may be a reference to a past mail or the utilize of terms like confirm or charge [1].

3. URL-based characteristics: These Properties look at when an IP address is utilized rather than a space title, the consideration of @ in joins, and number of cycles in joins and so on[1].

4. Script-based features: These features hunt for JavaScript, Pop-up window code, on-click exercises and other script- based highlights within the mail [1].

5. Sender-based characteristics: These characteristics give data around the sender such as the contrast between the sender’s address and the response to the address[1].

**Table 1: Comparative study of Phishing Emails**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No** | **Paper Name** | **Algorithm Used/Technology** | **Dataset** |
| 1 | Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey | Support Vector Machine (SVM) algorithm | Spam Assassin Public Corpus and the Nazario Phishing Corpus datasets |
| 2 | A Systematic Literature Review on Phishing  Email Detection Using Natural Language  Processing Techniques | Support Vector Machine (SVM)  K-Nearest Neighbours (KNN) | Enron dataset |
| 3 | Using Syntactic Features for Phishing Detection | Phishing Detection email | Nazario Phishing Corpus datasets |

Table 1 shows the comparative study of Phishing Emails we refer the survey of different papers. In this survey mostly used algorithms are the SVM (Support Vector Machine) algorithm [1-3]. The analysis of phishing emails are done on most popular dataset Nazario Phishing Corpus [1,3,4]. It is firstly used the rule-based phishing email detection approach [7].

1. **Speech Recognition:**

In NLP the specific speech recognition algorithm is used which focus on interaction between human and machine through language, speech and text. In 1952 at bell laboratories designed the “Audrey” system which could recognize single word speaking digits aloud. Later in the year of 1960 computer scientist have been researching ways & means to create computer record, clarify and recognize the human speech. The first speech recognition systems were focused on numbers, not words. The speech recognition are categorize into different technique like HMM (Hidden Markov Model) & GMM (Gaussian Mixture Model), Machine Learning, Speaker normalization algorithms, speech recognition algorithms, selection algorithm, Convulsion Neural Network.

Basically the hidden models are composed of hidden variables and observable.

In the survey of speech recognition algorithm mostly used HMM (Hidden Markov Model) [9-12] algorithm. In paper [9] HMM algorithm is used for the discrimination and robustness issue for speech recognition. HMM algorithm computationally efficient when it is compared with the Viterbi algorithm.

**Table 2: Comparative study of Speech Recognition**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No** | **Paper Name** | **Algorithm Used/Technology** | **Dataset** |
| 1 | The Kaldi Speech Recognition Toolkit | Gaussian mixture models (GMM),  Hidden Markov Model (HMM) | Linguistic Data Consortium (LDC) |
| 2 | A Review on speech recognition technique | Means algorithm  GMM Model  HMM Model  Sunspace Projection algorithm  K- Means algorithm | Feature extraction technique used. |
| 3 | Hidden Markov Models for Speech Recognition | Hill-climbing algorithm  Baum-Welch algorithm  k-means algorithm  Viterbi algorithm | - |
| 4 | Automatic Speech Recognition – A Brief History of the Technology  Development | Beam search algorithm  Speech clustering algorithms  Baum-Welch algorithm | - |
| 5 | Automatic Speech Recognition and Speech Variability: a  review | Simple peak counting algorithm  Phoneme-dependent frequency  warping algorithms  Speech recognition algorithms | - |
| 6 | Automatic speech recognition: a survey | Support vector machines Algorithm, | Libri Speech,  CHiME-5,  TED-LIUM Corpus |

**III ALGORITHMS**

NLP branches use various algorithms for mining the data, recognition of data, and prediction of data. Here table 3 shows the details of algorithms used in the various branches.

**Table 3: Use of algorithm of various Natural Language Processing branches**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Algorithm Used** | **Natural Language Processing Branches** |
| 1 | Support Vector Machine (SVM) algorithm | Email Mining, Text Mining |
| 2 | K-Nearest Neighbours (KNN) | Email Mining |
| 3 | K- Means Algorithm | Speech Recognition |
| 4 | Gaussian mixture Models (GMM) | Speech Recognition |
| 5 | Hidden Markov Model (HMM) | Speech Recognition |
| 6 | Baum Welch Algorithm | Speech Recognition, Email Mining |
| 7 | Hill- Climbing Algorithm | Speech Recognition |

1. **MACHINE LEARNING ALGORITHM**

Algorithms and statistical models that are programmed to learn from data, therefore accepting and inferring design within them. This enables computers to perform specific tasks without explicit command from a human operator [19]. Machine Learning Algorithm basically categorize in to two types such as i) Supervised Learning Algorithm ii) Unsupervised Learning Algorithm.

1. **SUPERVISED LEARNING ALGORITHM**

Supervised learning Algorithm refers to machine learning tasks whereby the goal is to recognize a function that best plan a set of inputs (e.g. image) to their correct output (label). It is based learning or training on prematched pairs. This is in contrast to unsupervised learning, where novel patterns such as groups or ‘clusters’ are identified in data without influence from prior knowledge or labeling[19].

1. **SUPPORT VECTOR MACHINE ALGORITHM**

The SVM is one of the most popular supervised learning algorithm. SVM algorithm was invented by Vladimir N. Vapnik in 1964 which could classify linear data. The task completed with the help of drawing hyper-plane between the data. There are two types of classifiers, linear classifier and non-linearclassifier[34].

1. Linear Classifier: Linear classifiers can classify the data by drawing the straight hyperplane such that maximum distance from either class data points. It is a parametric model, and it separates the data faster.

2. Nonlinear Classifier: Straight hyperplane concept can’t be used everywhere as data is dispersed. To handle this problem, authors suggested the kernel which converts two-dimensional space into higher dimensional space. The inner product of data values with kernel function is used to separate the data. With SVM mostly polynomial kernel and Gaussian (radial basis function) kernels are used.

a. Gaussian (Radial Basis Function) Kernel: It is used when the data is not linearly separable. If the size of training set increased, the complexity of the kernel also increased. It uplifts the samples into higher dimensional feature space which is separated by the linear boundary.The kernel equation is written as follows

K(xi,xj)=exp( - 𝜸 ⃦xi-xj  ⃦2 ) where 𝜸>0

b. Polynomial Kernel: It is the parametric model. It determines the similarity of input features and their combinations. Even though the degree ‘d’ of this function is high, it solves the problem only with multiplications.

K(xi,xj)=( 𝜸 xiT xj + r )d  where 𝜸>0

In earlier days, the SVM was used for binary classification, and afterwards, they can solve the multi-class problems.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

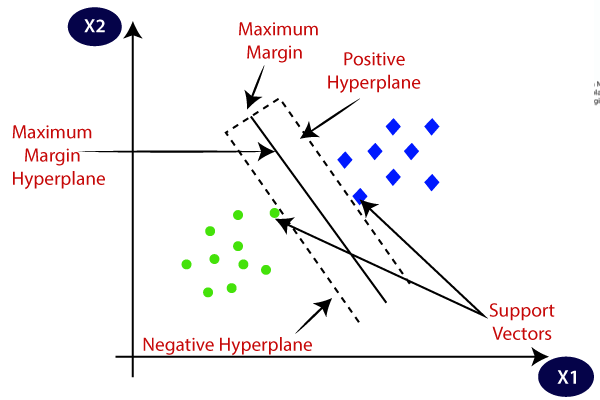


Fig: 1 Support Vector Machine

1. **HILL- CLIMBING ALGORITHM**

Hill climbing algorithm is a local search algorithm which continuously moves in the direction of increasing elevation/value to find the peak of the mountain or best solution to the problem. It terminates when it reaches a peak value where no neighbor has a higher value. Hill climbing algorithm is a technique which is used for optimizing the mathematical problems. [21]. Researcher used a global color histogram to find the peaks, step 5 takes into account the spatial information of the pixels when forming the segments, i.e. only spatially close (neighboring) pixels that lead to the same peak are grouped into one segment [34].

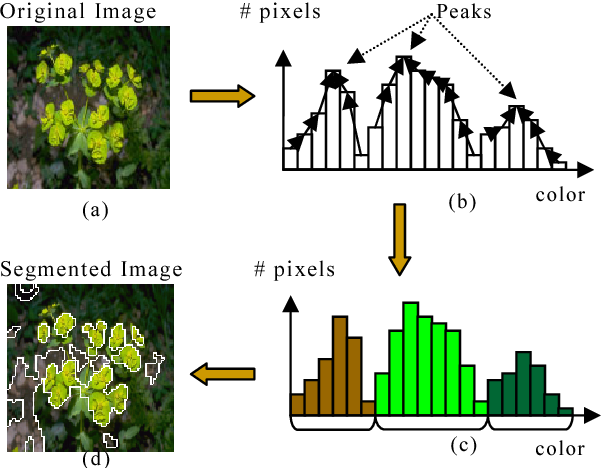


Fig. 2 An example showing the segmentation process: (a) original image, (b) hill-climbing process on the one-dimensional color histogram (hue component only), (c) the result of the hill climbing process, where 3 clusters are identified, and (d) the segmented image.[34]

1. **k-NEAREST NEIGHBOUR ALGORITHM**

k-NN was first introduced by E. Fix and J. Hodges in 1951. It is a nonparametric approach for the classification of data. The k-closest samples in the feature space used to train the data. All the computations required for the classification differs until the last step of classification. Every neighbour of the point has weight. It is the distance of the neighbours from that point. The results of the experiment vary with the value of ‘k’ [21]. **The data-point is classified on the basis of its k Nearest Neighbors, followed by the majority vote of those nearest neighbors; a query point is assigned the data class which has the most representatives within the nearest neighbors of the point.**

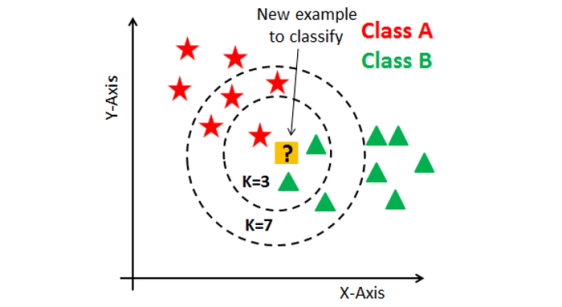


Fig 3: K- Nearest Neighbour Algorithm

1. **UNSUPERVISED LEARNING ALGORITHM**

Unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Rather, models itself find the hidden patterns and understanding from the given data. It can be compared to learning which takes place in the human brain while learning new things.

1. **K-MEANS ALGORITHM**

The simplest and most popular among iterative and hill climbing clustering algorithms is the K-means algorithm (KMA). This algorithm may converge to a suboptimal split-up. Since stochastic optimization approaches are good at avoiding convergence to a locally optimal solution, these approaches could be used to find a globally optimal solution. The stochastic approaches used in clustering include those based on simulated annealing, genetic algorithms, evolution strategies and evolutionary programming. Typically, these stochastic approaches take a large amount of time to converge to a globally optimal partition. [22].

The below diagram explains the working of the K-means Clustering Algorithm:

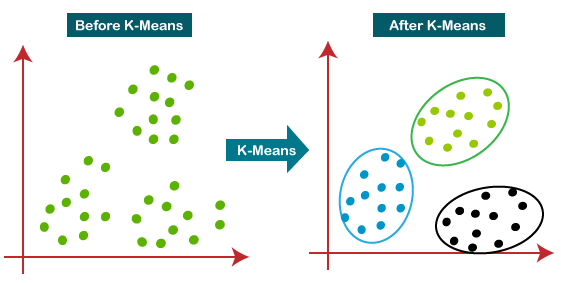


Fig 4 K-Means Algorithm

1. **BAUM-WELCH ALGORITHM**

Baum-Welch algorithm uses Lagrangian relaxation on a convert optimization problem, and we optimize the parameters for the dual problem which itself is a relaxed problem [23].

1. **HIDDEN MARKOV MODEL**

There are several methods for capturing speech and commonly used functions for the HMM and is a “parametric probability density function represented as a weighted sum of Gaussian component densities”. The HMM uses the Gaussian component parameters as the base classifier and acquires the temporal variations while the GMM captures the special variations. This allows it to efficiently and effectively handle time-sequences [24]. HMMs, the number of states varies depending on the length of the sequences in the family. The layout is strictly repetitive: In a profile HMM, there are three types of states: Match, Insert, and Delete. Match states represent informative positions in a family, while insert states represent a position. every step, the process moves from left to right, to the next distinct residue position, until the end state has been reached[36].

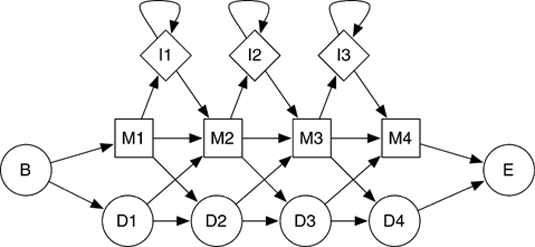


Fig 5

Fig: Schematic representation of a profile hidden Markov model of length 4. The states are begin (B), end (E), match (M), insert (I), and delete (D). Rectangular states (M and I) emit amino acids or nucleotides, round states are silent. A sequence family is represented by a characteristic distribution of amino-acid emission probabilities at every match or insert state. Delete states are equivalent to gaps.

1. **GAUSSIAN MIXTURE MODEL**

GMMs are widely used as probability distribution features, such as vocal-tract related spectral features in a speaker recognition or emotion recognition systems. GMMs having advantage that are more appropriate and efficient for speech emotion recognition using spectral feature of speech [24].

**IV CONCLUSION**

The paper explored the various different Natural Language Processing areas like Text Mining, Email-Mining, and Speech Recognition. Text Mining used support vector machine algorithm, K-means Algorithm, Clustering Algorithm. Now a days all official work is done through email. These emails are also used for various malicious activities. Hence many authors mined emails to find phishing, spam and detection of emails from the important emails. Emails are also used to recognize the sentiments of the persons which are used in the company for various activities like promotion, motivation. Speech Recognition fastens the work of every human being by listening commands. Various devices which are available in the market are Bluetooth, Alexa. NLP is using different algorithms like HMM (Hidden Markov Model) & GMM (Gaussian Mixture Model),Support Vector Machine(SVM),K-Means Algorithm, Baum-Welch algorithm, speech recognition algorithms, beam search algorithm, phoneme-dependent frequency warping algorithms, Hill-climbing algorithm, speech clustering algorithms, Means algorithm, Sunspace Projection algorithm for taking decisions and recognizing a data.

**REFERENCES**

1. Bogawar PS, Bhoyar KK. Email mining: a review. International Journal of Computer Science Issues(IJCSI). 2012 Jan;9(1).
2. Bogawar PS, Bhoyar KK. Soft computing approaches to classification of emails for sentiment analysis. InProceedings of the international conference on informatics and analytics 2016 Aug 25 (pp. 1-7).
3. Bogawar PS, Bhoyar KK. Comparative study of classification approaches for e-mail analysis. International Journal of Information and Computer Security. 2020;13(3-4):411-27.
4. Khonji M, Iraqi Y, Jones A. Phishing detection: a literature survey. IEEE Communications Surveys & Tutorials. 2013 Apr 15;15(4):2091-121.
5. Aleroud A, Zhou L. Phishing environments, techniques, and countermeasures: A survey. Computers & Security. 2017 Jul 1;68:160-96.
6. Fette I, Sadeh N, Tomasic A. Learning to detect phishing emails. InProceedings of the 16th international conference on World Wide Web 2007 May 8 (pp. 649-656).
7. Hamilton K, Nayak A, Božić B, Longo L. Is neuro-symbolic AI meeting its promises in natural language processing? A structured review. Semantic Web. 2022 Feb 24(Preprint):1-42.
8. Ferrone L, Zanzotto FM. Symbolic, distributed, and distributional representations for natural language processing in the era of deep learning: A survey. Frontiers in Robotics and AI. 2020 Jan 21;6:153.
9. Pacheco ML, Roy S, Goldwasser D. Hands-On Interactive Neuro-Symbolic NLP with DRaiL. InProceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations 2022 Dec (pp. 371-378).
10. Muthukumar P, Zhong J. A stochastic time series model for predicting financial trends using nlp. arXiv preprint arXiv:2102.01290. 2021 Feb 2.
11. Tan AH. Text mining: The state of the art and the challenges. InProceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases 1999 Apr 26 (Vol. 8, pp. 65-70).
12. Hotho A, Nürnberger A, Paaß G. A brief survey of text mining. Journal for Language Technology and Computational Linguistics. 2005 Jul 1;20(1):19-62.
13. Hearst M. What is text mining. SIMS, UC Berkeley. 2003 Oct 17;5.
14. Li Y, Algarni A, Albathan M, Shen Y, Bijaksana MA. Relevance feature discovery for text mining. IEEE Transactions on Knowledge and Data Engineering. 2014 Nov 24;27(6):1656-69.
15. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, Silovsky J. The Kaldi speech recognition toolkit. InIEEE 2011 workshop on automatic speech recognition and understanding 2011 (No. CONF). IEEE Signal Processing Society.
16. Gaikwad SK, Gawali BW, Yannawar P. A review on speech recognition technique. International Journal of Computer Applications. 2010 Nov;10(3):16-24.
17. Juang BH, Rabiner LR. Automatic speech recognition–a brief history of the technology development. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara. 2005 Jan;1:67.
18. Benzeghiba M, De Mori R, Deroo O, Dupont S, Erbes T, Jouvet D, Fissore L, Laface P, Mertins A, Ris C, Rose R. Automatic speech recognition and speech variability: A review. Speech communication. 2007 Oct 1;49(10-11):763-86.
19. Schank RC. What is AI, anyway?. AI magazine. 1987 Dec 15;8(4):59-.
20. Du‐Harpur X, Watt FM, Luscombe NM, Lynch MD. What is AI? Applications of artificial intelligence to dermatology. British Journal of Dermatology. 2020 Sep 1;183(3):423-30.
21. Beck J, Stern M, Haugsjaa E. Applications of AI in Education. XRDS: Crossroads, The ACM Magazine for Students. 1996 Sep 1;3(1):11-5.
22. Buchanan BG. A (very) brief history of artificial intelligence. Ai Magazine. 2005 Dec 15;26(4):53-.
23. Du‐Harpur X, Watt FM, Luscombe NM, Lynch MD. What is AI? Applications of artificial intelligence to dermatology. British Journal of Dermatology. 2020 Sep 1;183(3):423-30.
24. Suthaharan S, Suthaharan S. Support vector machine. Machine learning models and algorithms for big data classification: thinking with examples for effective learning. 2016:207-35.
25. Ohashi T, Aghbari Z, Makinouchi A. Hill-climbing algorithm for efficient color-based image segmentation. InIASTED International Conference on Signal Processing, Pattern Recognition, and Applications 2003 Jun 30 (pp. 17-22).
26. Krishna K, Murty MN. Genetic K-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 1999 Jun;29(3):433-9.
27. Hsiao R, Tam YC, Schultz T. Generalized baum-welch algorithm for discriminative training on large vocabulary continuous speech recognition system. In2009 IEEE International Conference on Acoustics, Speech and Signal Processing 2009 Apr 19 (pp. 3769-3772). IEEE.
28. Ault SV, Perez RJ, Kimble CA, Wang J. On speech recognition algorithms. International Journal of Machine Learning and Computing. 2018 Dec;8(6):518-23.
29. Hassani H, Silva ES, Unger S, TajMazinani M, Mac Feely S. Artificial intelligence (AI) or intelligence augmentation (IA): what is the future?. Ai. 2020 Apr 12;1(2):8.
30. Zhong H, Xiao C, Tu C, Zhang T, Liu Z, Sun M. How does NLP benefit legal system: A summary of legal artificial intelligence. arXiv preprint arXiv:2004.12158. 2020 Apr 25.
31. Chowdhary K, Chowdhary KR. Natural language processing. Fundamentals of artificial intelligence. 2020:603-49.
32. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. Journal of the American Medical Informatics Association. 2011 Sep 1;18(5):544-51.
33. Raina V, Krishnamurthy S, Raina V, Krishnamurthy S. Natural language processing. Building an Effective Data Science Practice: A Framework to Bootstrap and Manage a Successful Data Science Practice. 2022:63-73.
34. Bogawar PS, Bhoyar KK. Comparative study of classification approaches for e-mail analysis. International Journal of Information and Computer Security. 2020;13(3-4):411-27.
35. Ohashi T, Aghbari Z, Makinouchi A. Hill-climbing algorithm for efficient color-based image segmentation. InIASTED International Conference on Signal Processing, Pattern Recognition, and Applications 2003 Jun 30 (pp. 17-22).
36. Beal M, Ghahramani Z, Rasmussen C. The infinite hidden Markov model. Advances in neural information processing systems. 2001;14.