iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 9 **Issue:** VII **Month of publication:** July 2021

**DOI:** https://doi.org/10.22214/ijraset.2021.37196

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 9 Issue VII July 2021- Available at www.ijraset.com*

# Automated Brain Tumor Prediction System using Natural Language Processing (NLP)

Gourav Sharma[1], Heena Khera[2], Nikita[3], Hanifa Arfi[4]
[1]*Department of Information Technology, Dr. Akhilesh Das Gupta Institute of Technology and Management, Delhi, India.*
[2, 3]*Department of Computer Science, Galgotias University, Noida,UP.*
[4]*Department of Information Technology, Birla Institute of Technology - Mesra, Ranchi, Jharkhand, India.*

***Abstract:** In this paper, we proposed an Automated Brain Tumor Prediction System which predicts Brain Tumor through symptoms in several diseases using Natural Language Processing (NLP). Term Frequency Inverse Document Frequency (TF-IDF) is used for calculating term weighting of terms on different disease's symptoms. Cosine Similarity Measure and Euclidean Distance are used for calculating angular and linear distance respectively between diseases and symptoms for getting ranking of the Brain Tumor in the ranked diseases. A novel mathematical strategy is used here for predicting chance of Brain Tumor through symptoms in several diseases. According to the proposed novel mathematical strategy, the chance of the Brain Tumor is proportional to the obtained similarity value of the Brain Tumor when symptoms are queried and inversely proportional to the rank of the Brain Tumor in several diseases and the maximum similarity value of the Brain Tumor, where all symptoms of Brain Tumor are present.*

***Keywords:** Brain Tumor, TF-IDF, Natural Language Processing (NLP), Cosine Similarity, Euclidean Distance.*

## I. INTRODUCTION

In this paper, Brain Tumor is predicted through queried symptoms, which is done by ranking the Brain Tumor through symptoms in several diseases then estimate chance of Brain Tumor. Rank of the Brain Tumor through queried symptoms in several diseases is calculated by Natural Language Processing (NLP) for queried symptoms [4] text processing, TF-IDF for term weighting, Inverted Index is used for effective use of storage with efficient information retrieval in real-time, Cosine Similarity Measure and Euclidean Distance are used for calculating angular and linear distance respectively between diseases and symptoms. Counting-Sort is used for sorting diseases on the bases of their rank on the basis of angular and linear distance. And the chance of the Brain Tumor is proportional to the obtained similarity value of the Brain Tumor when symptoms are queried and inversely proportional to the Rank of the Brain Tumor in several diseases and the Maximum Similarity value of the Brain Tumor when all symptoms of Brain Tumor are present.

Here, diseases and documents, disease's symptoms and document are interchangeable respectively. Before we are going through the proposed technique, results and conclusion, we are elaborating the term Brain Tumor.

Brain Tumor a brain tumor is an abnormal collection of cells in the brain. The skull is very rigid and the brain is enclosed, so any growth inside such a restricted space can cause problems. Brain tumors can be cancerous (malignant) or non-cancerous (benign) Brain Tumor don't depends upon Age and Sex. Symptoms of Brain Tumor are following: (i) Headaches, Nausea (ii) Vomiting (iii) Changes in speech, vision, hearing (iv)Problem in balancing, walking, (v) Changes in mood, personality, (vi) Muscle jerking (vii) Tingling in arms, legs.

## II. PROPOSED TECHNIQUE

The proposed technique is divided into two parts:

*1)* Database Creation Process: (i) Documents Text Processing using NLP, (ii) Documents Term Weighting and Magnitude Calculation and (iii) Create Database & Inverted Index[8][9]

*2)* Prediction of Brain Tumor: (i) Queried Symptoms Text Processing using NLP, (ii) Queried Symptoms Term Weighting and Magnitude Calculation, (iii) Retrieve Documents Term Weighting and Magnitude from Database and Inverted Index[10][11][12](iv) Evaluate Rank and Similarity value of Brain Tumor through symptoms, and (v) Estimate Chance of Brain Tumor

### *A. Database Creation Process*

*1)* Parsing1 of Disease's Symptoms

*2)* Remove Stop-words2

*3)* Stemming[3] (Optional)

*4)* Calculate TF-IDF and Square of Magnitude of Documents

Let, t is a term which is exists in DFt documents and

N = Total Number of Documents

TFd, t = how many times a term (unique word) occurs in a document (assumes d)

Nd= Total number of terms in d

Normalized TFd, t = TFd, t / Nd

IDFt = 1 + log10(N/df)

TF-IDFd, t = Normalized TFd, t * IDFt

Let Dn n, Tm stands for term m and D1 for Brain Tumor.

Calculate Square of Magnitude of Documents and Save IDFt and Magnitude of Documents in Database

$|Dd| = \sqrt{\sum} (TF - IDFd,t)2$ Or,

$|Dd|2 = \sum (TF - IDFd,t)2$

*5)* Create Document Vs TF-IDF Inverted Index and Save Square of Magnitude of Documents in Database.

Note: Diseases / documents and terms are to entertain which are not matches with at least one term of Brain Tumor / Document.

*B. Prediction of Brain Tumor*

*1)* Parsing of Queried Symptoms

*2)* Remove Stop-words

*3)* Stemming (Optional)

*4)* Calculate TF-IDF[2] and Magnitude of the Queried Symptoms Same formulas have been used for TF-IDF and Magnitude of the Queried Symptoms calculation which is used in database Creation Process in section I (d). Let Q stands for query and Ti stands for term i.

*5)* Get TF-IDF of queried terms in Documents from Inverted Index4 and Magnitude of Documents from Database

*6)* Calculate Cosine Similarity between Query and Documents

Cosine Similarity[5]

(Q, Dd) = Dot product (Q, Di) / (|Q| * | Dd|)

Where, Q and Dd are two vectors

Sort Diseases on the bases of Cosine Similarity

*7)* Calculate Euclidean Distance[6] between Query and Documents

Euclidean Distance

$(Q, Di) = \sqrt{|Dd|2 + |Q|2 - 2\sum (Qj \times Dd,j)}$ mj=1

Or,

*a)* Parsing [1] is a process of extracting unique words from documents, with words in a same case (Full Lower or Full Upper Case) without redundant data like symbols etc. from raw data.

*b)* Stop-Words [1][5][6] are those words which are commonly used in various sentences, short term words, such as “the, is, at, which, in, on”. Stop-words are generally removed for better ranking of documents.

*c)* Stemming [1] is a process for reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. Some time Stemming give poor result so, we have taken Stemming as an optional feature in the proposed technique.

*d)* Inverted List [1] is an inverted file is a word-oriented mechanism for indexing a text collection in order to speed up the searching task. Inverted List contains two things (a) “Vocabulary is the set of all distinct words in the text” and (a) “Occurrences lists containing all information necessary for each word of the vocabulary”.

*e)* Cosine Similarity [3][2] is used for calculating the similarity or angular distance between any two vector points or two documents or document and query, which are in vector form i.e. Cosine Similarity (v1, v2) = Dot product (v1, v2) / (|v1| * |v2|); where v1 and v2 are two vectors.

*f)* Euclidean Distance [1] is used for calculating the separation or linear distance between any two vector points or two documents or document and query, which are in vector form i.e.

$d(x,y) = \sqrt{\sum} (xi - yi)2\ pi=1$

$= \sqrt{|Dd|2 + \sum ((Qj - Dd,j)2 - Dd,j\ 2\ )\ mj=1}$

Where,j=1 d,j

Qj is jth matched term in TF-IDFq, t

Dd, j is jth matched term in TF-IDFd, t of d-th document

8) Estimate Rank and Sort Documents on the basis of Rank

9) Sort Diseases using counting sort on the bases of Rank

10) Estimate Chance of Brain Tumor [3][9][8]

Chance of Brain Tumor = (y/x)*(R/(R+C*(R-1))))*100%

And, If R=0 then Chance of Brain Tumor ignored (0%). Where,

y = Obtained Cosine Similarity value in query when symptoms are queried

x = Maximum Cosine Similarity value of Brain Tumor when all symptoms of Brain Tumor are present.

C = Coefficient

R = Ranking of Brain Tumor in several diseases through symptoms Let, x = 1 (for Cosine Similarity) and C = 1

i. If R = 1 and y = 1 then

Chance of Brain Tumor = (1/1)*(1/1)*100% = 100%

ii. If R = 1 and y = 0.812751 then

Chance of Brain Tumor = (0.812751/1)*(1/1)*100% = 81.2751%

iii. If R = 8 and y = 0.219781 then

Chance of Brain Tumor = (0.219781/1)*(8/8)*100% = 21.9781%

iv. If R = 0 and y = 0.0 then Chance of Brain Tumor = 0%

Note: Chance of Brain Tumor is also depends upon Number of Documents in Database because IDF is depended on Number of Documents.

11) Show Result in Human Readable Form

Display Diseases ID, Name and Symptoms according to their ranking.

## III. RESULT

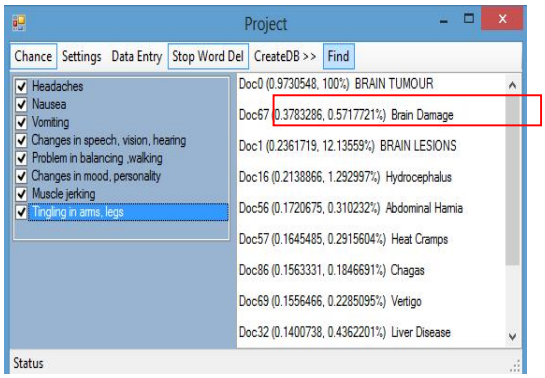



Fig. 1.Ranking of Brain Tumor in several diseases against different combination of symptoms.

Fig. 2.

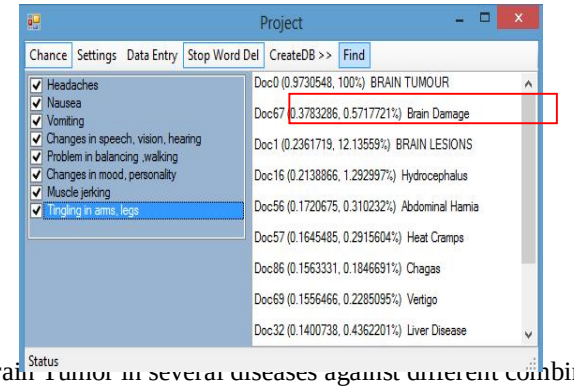


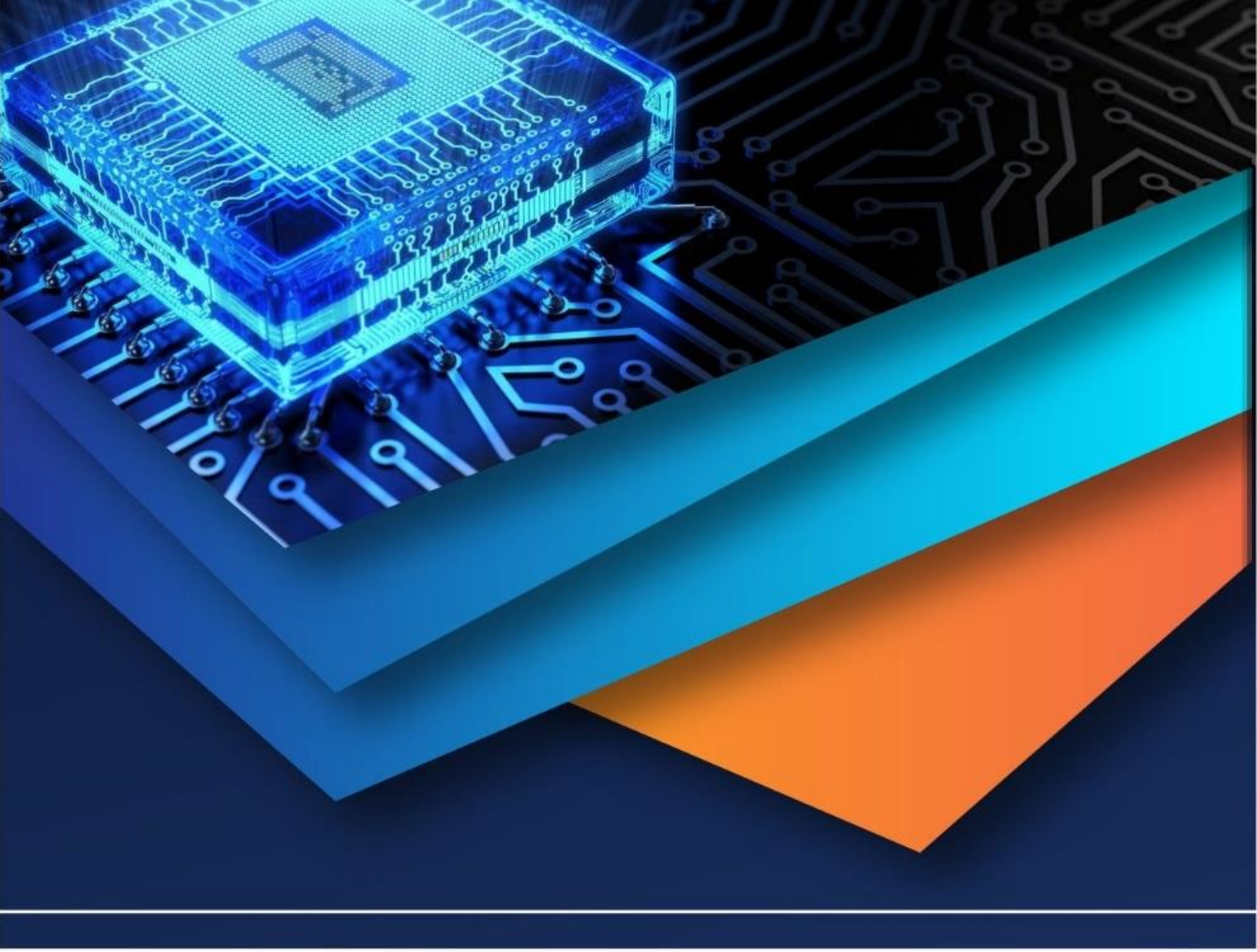

Fig. 3.Chance of Brain Tumor in several diseases against different combination of symptoms.

## IV. CONCLUSION

In this paper, we are predicting Brain Tumor using Brain Tumor Ranking in several diseases by symptoms and novel mathematical strategy. The developed Brain Tumor decision support system can be used by physician to automatically diagnose the Brain Tumor by entering basic sign and symptoms of a patient. The proposed system is not only used for diagnosis but also be used to store and view the result of diagnosis for further reference. Here, TF-IDF[7] is used for calculating term weighting for better result based on query observation. Conventionally only angular distance else only linear distance is used for ranking of documents. But, the proposed paper both Cosine Similarity and Euclidean Distance are used for Brain Tumor ranking through symptoms in several diseases for better ranking. Counting-Sort is used for sorting document on the bases of their rank. Inverted File Indexing is used for effective use of storage with efficient ranking in real-time. Novel mathematical strategy is used for calculating chance of Brain Tumor in percentage.

## REFERENCES

[1] Nikita, Yaseer Ali Ahmad and G.Sahoo: Ranking of Brain Tumor through Symptoms in Several Diseases. In: International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.55 (2015), Page No. 2994-2999, Research India Publications. http://www.ripublication.com/ijaer.htm
[2] H. Wu, R. Luk, K. Wong and K. Kwok: Interpreting TF-IDF term weights as making relevance decisions. In: TOIS, vol. 26, no. 3, pp. 1-37, 2008.
[3] S.E. Robertson: The Probability Ranking Principle. In: Ir, Journal of Documentation, Vol. 33 Iss: 4, pp.294 – 304, 1977.
[4] Vanisree K and Jyothi Singaraju: Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptom using Neural Networks. In: International Journal of Computer Applications, (0975 – 8887) Volume 19– No.6, April 2011
[5] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze: An Introduction to Information Retrieval. Cambridge, England: Cambridge University Press, April, 2009.
[6] G. Salton and M. J. McGill: Introduction to modern information retrieval. 1983.
[7] Yaseer Ali Ahmad and Dr. Abhijit Mustafi: Document Ranking Strategy Enhancement using Paragraph Ranking. In: International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.55 (2015), Page No. 2973-2978, Research India Publications. http://www.ripublication.com/ijaer.htm
[8] Bast, H., Buchhold, B.: An index for efficient semantic full-text search. In: 22nd ACM Int. Conf. on CIKM, pp. 369– 378 (2013)
[9] Burton-Jones, A., Storey, V.C., Sugumaran, V., Purao, S.: A heuristic-based methodology for semantic augmentation of user queries on the web. In: Song, I.-Y., Liddle, S.W., Ling, T.-W., Scheuermann, P. (eds.) ER 2003. LNCS, vol. 2813, pp. 476–489. Springer, Heidelberg (2003)
[10] Carpineto, C., et al.: Improving retrieval feedback with multiple term-ranking function combination. ACM Trans. Inf. Syst. 20(3), 259–290 (2002)
[11] Chandramouli, K., et al.: Query refinement and user relevance feedback for contextualized image retrieval. In: 5th International Conference on Visual Information Engineering, pp. 453–458 (2008)G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY