Journal of Engineering, Electrical and Informatics
Vol.2, No.2 Juni 2022
e-ISSN: 2809-8706; p-ISSN: 2810-0557, Hal 23-37

DEEP LEARNING ALGORITHM FOR CONNECTING SCIENTIFIC RECORDS AND SOCIAL PLATFORM

Budi Santoso
Universitas Sains dan Teknologi Komputer

Agustinus Budi Santoso
Universitas Sains dan Teknologi Komputer

Eko Siswanto
Universitas Sains dan Teknologi Komputer

Jl. Majapahit 605, Semarang, telp/fax : (024) 6723456

Abstract*. In the healthcare industry, professionals develop big amounts of disorganized data. The complexity of this data and the loss of computational capability lead to delays in the investigation. Nevertheless, with the advent of Deep Learning algorithms and connection to computing power such as Graphic Processor Units (GPUs) and Tensor Processing Units (TPUs), text and image processing has become usable. Deep Learning (DL) data bring about a big outcome in Natural Language Processing (NLP) and computer perception. The main purpose of this study is to build an undivided approach that can relate social platforms, literature, and scientific records to develop an approach to medicinal education for the public and experts.*

This study focuses on NLP in the healthcare industry and compiles data by Electronic Medical Records (EMR), medical literature, and social platforms. The framework proposed in this study is one for connecting social platforms, medical literature, and Electronic Medical Records scientific records using Deep Learning algorithms. Linking data sources requires defining the relationships between them, and finding concepts in medical texts. The National Library of Medicine (NLM) introduced the Unified Medical Language System (UMLS) and uses this system as the basis for the proposed system. The dynamic nature of a social platform can be recognized and supervised methodologies can be applied under supervision to develop conception. Named entity Recognition (NER) enables the active eradication of data or individuals by the pharmaceutical literature.

The model for processing EMR scientific records was extended through transfer learning. The result includes a unified, end-to-end, web-based systems solution that brings together a social platform, literature, and scientific records, and enhances access to medical knowledge for the public and experts. As a result of this research, an integrated model that can link social platforms, literature, and scientific records is proposed to prove useful in increasing access to medical knowledge for the public and experts and being able to meet knowledge relational needs. This study makes a lot of improvements, along with data collection, the introduction of two databases, the creation of a new noise elimination scheme for the social platform, the building of a Deep Learning model for titled individual acceptance, and the adoption of transfer learning to establish the model.

Keywords*: Social Platform, Natural Language Processing, Deep Learning, Tensor Processing Unit, Electronic Medical Record*

Received April 07, 2022; Revised Mei 2, 2022; Juni 22, 2022
* Budi Santoso

INTRODUCTION

The aspect of hospital services necessarily changes people's lives. Hospital experts take each duty, or hospital affair, and calculate remote to build the aspects of their services. Today, enormous repository material and complicated instruction administration electronics grant the particular expert to stock, measure, and record every hospital affair. The files on this material develop the aspects of hospital services. Nevertheless, corresponding and mutating files in current healthcare systems has proven to be complicated, time-consuming, and expensive (Yu et al., 2012). Having an essential and entire files revolution lack a deep perceptive of file processing. The lack of data sincerity drawn in the lab and ICU results in limited data accuracy. Electronic health records (EHR) and Electronic medical records (EMR) were the first pioneering systems to collect data quickly, and expertly and overcome the limitations of older data collection methods.

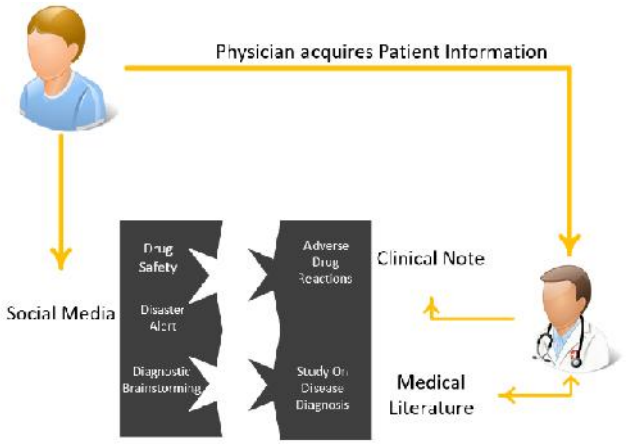

Figure 1. Definition of the difference between records made by physicians in the scientific record form and biomedical and patient publications on the social platform

Healthcare texts are categorized into biomedical publications, scientific records, and social commentaries. The biomedical publication collects the texts of Medical Doctors (MDs) and combines them with those of alternative physicists. MDs, therapists, nurses, radiologists, etc. produce a variety of hospital sign and precept physicians' experience and practices. Social platform texts can be generic conservation, suggestions, or particular experience on healthcare topics. Even if they are not experts in their field, public contributions are still important to research. Social network resources represent public health beliefs and help understand many topics such as diagnoses, drugs, and claims.

Suggested Solution

The main objection to identification text in healthcare is a disorganized data repository and data improvement which is quite complicated. To cut the complication of recovering scientific or biomedical text data, it is also necessary to catch 4 different appearances in the content. The first is negations, sentences, or phrases that mention symptoms the patient does not have or report an unsuccessful diagnosis. The next is confidence, censure, or idiom that notices the possible problem. For the 3rd is humanity, the record of idom can modify more time while analysis. Humanity is separated into two parts: First, the historical component, when the patient had symptoms at least 14 days before the date of examination. Second, is the hypothetical breakdown, when the doctor

assumes that the patient can have the positive disorder for further examination. Studies show that distinguishing provisional statements complicates the identification of recent issues as long as there is a big probability of transitions between historical, instant, and hypothetical to present. Another reason for this complexity lies in the fact that the conclusion has small data to track on this transition (Mowery (2014; Chu et al., (2007)). For the 4th, family background, sentences or phrases that record the patient's family medical history. "The patient's father has a history of CHF (Congestive Heart Failure)." Having annotated the words with these 4 features, it becomes necessary to tackle the identical objection. Named Entity Recognition (N-ER) addresses this challenge by recognizing entities and matching them together. This entity will help catch the 4 aspects in medical texts and link health-related social platform domains with publications and scientific records. The main goal of this research is to build an undivided approach that can connect social platforms, literature, and scientific records to develop an approach to medical awareness for the public and experts.

METHODOLOGY

Latent Semantic Analysis (LSA)

The likelihood of discovery groups of words with the same context is expected in the compilation of annals on the same topic. LSA treasure trove is a class of texts that can make a group of archives. Scott Deerwester, (1990) proposed “a model for finding these words among archives”. Dumais, (2004) describes “that collecting frequencies sublinearly will increase results. Because archives and terms are in the same space vector, records can be defined with several words known as topics.

Latent Dirichlet Allocation (LDA) – LDA MAL-LET

“LDA is a 3-levels of hierarchical Bayesian model” (Jordan et al., (2003)). LDA aims to scale down the information dimensions and build topic-based representations of archives by a group of words that defines each topic (Jordan et al., (2003)). “The research team at UMASS AMHERST produced the MALLET toolkit” (McCallum, (2002). This toolkit is mutual and consists of a variety of advantageous affair design approaches, equally applying sample-based LDA.

Biotherm Topic Models

This model determines the issues of dispersion of information in short texts. Yan et al., (2013) proposed “the Biterm model, which bases word connectivity on patterns rather than documents”. The Biotherm model builds terms, such as "breast cancer" and "digital health care" by looking for the relatedness among the words in the entirety. The topic of this model is symbolized as “z”, the Dirichlet word circulation for different topics is “ϕ” and the topic distribution is “θ”. The different Biterm contains 2 words this is “w_i”, “w_j”), and then combined possibility is calculated as the probability of the entire corpus as follows:

$$P(B) = \prod_{(i,j)} \sum_{z} \theta \, \phi_{(i\backslash j)} \phi_{(j\backslash z)}$$

Support Vector Classification (SVC)

The continuous SVC 1 of the simpler analysis approach, creates a "best fit" edge. The different row designs are a result of bordered information by accustomed two divisions collapsing on the different sides of the line (Figure 4).

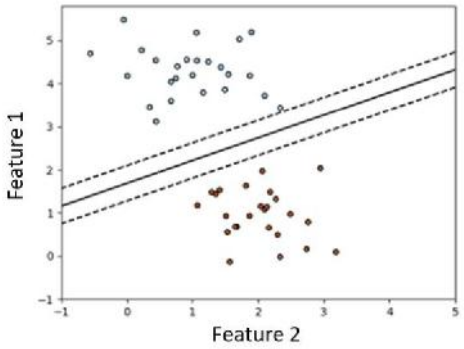

Figure 2. Illustrating result Boundaries for Analysis

Multinomial Naive Bayes

The multinomial Naive Bayesian classifier extends the traditional Naive Bayesian classifier, and it shows a simplified formulation and representation of each in Eq. below.

$$p(C_k|x) = \frac{p\,(C_k)p(x\backslash C_k)}{p(x)}$$

$$posterior = \frac{prior * likelihood}{evidence}$$

The main feature of Naive Bayesian probability is the assumption of independence of its features (hence Naive). Predicting the possibility by 100 comments of samples, X of them as "Yes", and set other as "No".

Text Enclosed

In structured information lists and lines are used to decide the data. Different lists decide the case, and the line serves as the appearance that describes the row. Appearances are continuous and simple to change for analytical models. Non-anatomical data equally text as an aspect is can't describe data preparation for numerical data as an objection. Various techniques equally Hash Vectorization, Term Frequency, Inverse Document Frequency (TF-IDF), GLOVE Stanford, and Google Word2Vec propose and establish text transformations. A sentence defines as a sequence of characters in files. The area of this feature is very important, so it can be concluded that D is represented by a feature vector $\vec{V}$. This vector needs to be converted to a vector of numbers representing Documents. Thus, all these techniques aim to represent text with vector numbers.

LITERATURE REVIEW

Social platform

Social networking sites serve as a platform for communication. "Twitter users share advice about health-related information" (Larson, (2010); Prier et al. (2011)). These authorities develop the national expectation in association and expand their compassion for diagnoses, drugs, and claims. "There are nearly 140 potential health uses of Twitter" (Terry,(2009)). "The most common uses are: disaster warning and response, diabetes management, Food, and Drug Administration drug safety alerts, biomedical device data capture and reporting, shift offering for nurses, and other healthcare professionals, diagnostic brainstorming, rare disease tracking, and resources connections, smoking cessation assistance, baby care tips for new parents and post-discharge.

Spatio-Temporal Analysis

Researchers use “Twitter data to detect expectation, danger investigation, and disclosure affair” (Hwang et al., (2013); Zhao et al., (2015); Kalantari, (2017)). The exceptionally valuable procedure on a social platform is the text starting with eyewitnesses. Such Tweets apply to a wide range of attitudes, equally administration, instruction, and experience systems in a variety of areas, including natural disasters, urban traffic, and healthcare. Part of the functional adoption of Twitter information is the recent disclosure of collapse, covering information, and compassionate public perceptions. Dictionary-based text analysis and Latent Dirichlet Allocation (LDA) is unsupervised topic modeling to provide journalists with better insights (Guo et al., (2016)). Hwang et al., (2013) suggested a climbable channel framework and database design collect and analyze data. Rathore et al., (2017) proposed “a Hadoop-based framework for performing real-time analysis of social platform data”. They generalize this scheme and catch workout feeding by essential emergency to health data. “Tagging recommendations and scientific topics identified in the document are other uses of Twitter data” (Nejdl et al., (2009)).

Deep Learning

Bidirectional Long Short-Term Memory

Detects the individual words or text depending on the body used. Treasure trove correlations among words by classical neural networks will not be effective due to the complexity of the sentences. “The architectural design of a recurrent neural network is an excellent choice for studying and analyzing sentence sequences. Recurrent neural networks are used in dialect design” (Mikolov et al., (2010); Vinyals et al., (2014); nevertheless, one of the deficiencies of adopting simple recurrent neural networks is the difficulty in capturing sequential long-term correlations. The long-term memory approach introduced to overcome the limitations of the recurrent neural network is a group of associated words. Applicable words will analyze the different individual sentences. To carry information in two directions this study uses the BiLSTM model.

Enclose layer

The enclosed layer builds a map among the words, a heavy numeric matrix in natural language processing to produce precise input to an artificial neural network. Layers of learned words and characters, word-level enclosed, characterizing different words as an aim, have to a dictionary of words collected by documents called V with size $|V|$, and D are as the size of the enclosed, which produce the size of the vector that describes the word. Accordingly, this research has $|V| * D$ for matrix size with different rows indicate to a single word that behaves like a discover table. At this level, “more define word-level enclose equally GLOVE from Stanford or Word2Vector from Google” (Manning et al., (2014; Mikolov et al., (2013)) is applied.

Studying the morphological characteristics of words such as suffixes and prefixes will add value to the analysis (Ma & Hovy, (2016); Santos & Zadrozny, (2014)). This study adopted the 2 kinds of aspect-level of enclosed (recurrent neural network and convolution neural network). In several issues, equally online learning needs to create a dictionary since the instruction can interfere. The target of contending is to cut down the area aspect of T-tokens. Patter necessarily outcome in lots of tokens "colliding" with each other because they are assigned to the same pattern. Since the inevitably various tokens break up, they will get an accurate aim depiction, and excessive approach by differentiating among them. Stew enclosed is adopted as a departure between the common word enclosed, and that built by a random stew function.

The layer of Conditional Random Field

Conditional Random Field (Pereira et al., (2001)) uses sequencing data segmentation and labeling. Research shows that "Conditional Random Field techniques at the sentence level accomplish improved outcome than single word investigation equally Maximum Entropy Markov Models and Hidden Markov Models" (Pereira et al, (2000); Ratnaparkhi, (1996)). Conditional Random Field mixed with Bidirectional Long Short Term Memory (Bi-LSTM) can improve the results of sequence analysis (Yu et al., (2015)). Figure 5 shows the sentence at the bottom of the figure gap over the LS-TM block, and the Conditional Random Field layer concludes the outcome during the individual entity type of word. The Conditional Random Field loss function is detailed as a scoring matrix, where A is a sentence formed as a sequence described by $[[\ x\ ]_1^T]$, then the function for calculating scores is $f\theta\ ([\ x\ ]_1^T)$. In the scoring matrix, Each element is defined by $[\ f_\theta\ ]_{i,\ t}$ Where it indexes the output tag and t addresses each word, this matrix represents a rating system that is independent of position.

$$s([x]_1^T, [i]_1^T, \theta) = \sum_{t=1}^{T} ([A]_{[i]_{t-1},[i]_t} + [f_\theta]_{[i]_t}, t)$$

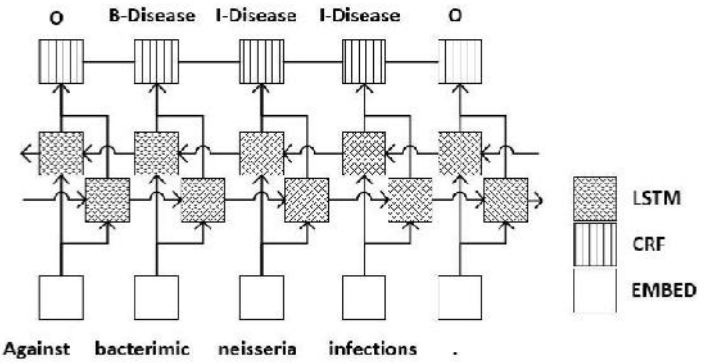

Figure 3: Demonstrating Sentence Transformation to outcome by combining the enclosed extension, two-way LS-TM, and Conditional Random Field

DATA AND METHODOLOGY

Public Datasets

A lot of standard openly accessible datasets as training and evaluating N-ER assignments. The 5 various databases were selected in the biomedical sphere focusing on the category to threat the approach used and appraise adaptation across domains. The BIO (Start, Inside, Outside) format is used to equip datasets and tag entities, preprocessing and splitting of all datasets into 3 files is also performed, it covers seventy percent, authorization consists of ten percent, and tests comprise twenty percent of the pool data. Ma & Hovy, (2016) described that the JNL-PBA dataset was driven by the GENIA corpus and annotated for use as a ground truth for detecting Proteins, RNA, Cell Lines, Cell Types, and DNA. BioCreative II Gene Mention (BC2GM) (Smith et al., (2008)) is a benchmark data set for training the N-ER task and has been adopted by a lot of departments to catch RNA labels consisting of BAN-NER, GLI-MI, and BIO-BERT that were built by Leaman & Gonzalez, (2008) and Oliveira et al., (2013). Mi & Thomas, (2009) described that "Pathway Curation is the main task in BioNLP 2013". This data set

is designed to handle case eradication in medical texts to support curation. Event extraction entities are identified first using this data set to catch the Genes or Gene Products, Complexes, Cellular Components, and Simple Chemicals before performing. Bergman et al., (2010) described about "The LINNAEUS dataset normalizes and recognizes species names mentioned in medical texts, the version in this study contains 100 random full-text documents converted to a stand-off format".

One of the prominent entities in medical texts is disease name, the NCBI disease test (Lu et al., (2014)) was chosen to evaluate the model in this study. This data set consists of seven hundred and ninety-three medical publication abstracts, including seven hundred ninety different disease names. Li et al., (2016) described that Bio-Creative V Chemical Disease Relation (BC5-CDR) combined 2 entities (chemistry and condition) and cited humans by thousand and fifty hundred Medical Publication papers.

Dataset	Type	Entity Frequency
JNLPBA	Gene/Protein	DNA: 8,392 protein: 27,032 cell type: 6,177 cell line: 3,380 RNA: 837
BIONLP13PC	Gene	gene: 5,399 complex: 719 cellular: 464 chemical: 1,155
BC2GM	Gene	gene: 15,047
LINNAEUS	Species	species: 2,103
NCBI-DISEASE	Disease	disease: 5,118
BC5CDR	Disease/Chem.	chemical: 5,185 disease: 4,098

Table 1. Types and density of individual Datasets (Summary)

Scientific records Data Collection

Medical Information Mart for Intensive Care (MI_MIC) vers-3 is an expanded database of electronic patient health records. MI-MIC data consist of laboratory test results, bedside critical hint analysis, the procedures anesthetic, caregiver notes, clarify reports, and deaths. Sha-Re/CLE-F e-Health generated datasets in 3 assignments to aid improved info comeback in natural language processing with scientific care approaches. Assignment 1 and assignment 2 concerned scientific record annotations, and the third task concerned website pages based on queries starting while learning clinical reports. One and assignment 1 and assignment 2 have twenty fine clinical words for discipline and one hundred clinical pieces of information for testing each. Assignment 1 consists of a glossary of distractions, and Assignment 2 consists of phrases/abstractions. Assignment 3 is a medical-associated website document set, 5 evolution objections, a website document outcome set, and fifty analysis queries (Suominen et al. (2013); Kelly et al. (2014Z)).

WELL-BEING TREND DISCLOSURE AND SPATIO-TEMPORAL INVESTIGATION SOCIAL PLATFORM

This research proposes a scheme for advising, clarifying, investigating, perceptive, and hunting the case in data streams. This study proposes a topic-tracking neural architecture that provides 83.34% accuracy, 83% precision, 84% recall, and 83.8% F-Score. This research introduces a model selector action by hybrid indicators to

overcome online topic disclosure and creates a mechanized data processing pipeline with 2 stages of cleaning. Routine and deep cleaning are practiced by various metaknowledge sources to improve data quality. Deep Learning and transfer learning techniques allocate health-related tweets by great certainly and better F-1S (Score). On this, the visualization uses to accept trending topics.

Keep track of social platform topics

The proposed model is made with five layers:

- Layer 1: data collected by Teewpy (python tool) (Kuperman et al., (2007)).
- Laye 2: consists of the cleaning and pre-processing methods for transforming tweets into processable vectors.
- Layer 3: the Word-2-Vec method that develops a matrix based on the vectors accepted by the last layer and uses the matrix to initialize a neural network to predict labeled tweets.
- Layer 4: the convolutional neural network classifier, where invisible tweets originating from Word2Vec are labeled.
- Layer 5: data that is not labeled to the LDA model is entered and a new topic is created.

Typically, sequence modeling is associated with a recurrent neural network. Nonetheless, the outcome shows various aspects. Bates et al., (2014) described “the Convolutional neural networks provide an important outcome in Natural Language Processing. Reddy & Aggarwal, (2015) and Chapman et al., (2011) apply the convolutional neural network for semantic parsing Shen 2014 uses it for query retrieval. Ibanda, (2009)) defines that Kalch -Banner 2014 uses it for sentence modeling and Bates et al., (2014) des scribed that Yoon Kim (2014) connects the Word-2-Vec model by the convolutional neural network. As part of this research project exploring classification models that can be used for adaptive systems. Aspect collection is one of the challenges, and Convolutional Neural Networks are applicable. The condition of the convolutional neural network is that they require a fixed input size; since tweets are limited to two hundred and eighty elements, filler for briefer tweets can be used, keeping the input size fixed. The research design consists of three convolutional layers with kernel sizes of 128-64-32 correspondingly. The system has a decrement of zero point five. The convolutional neural network model will update the Word2Vec problem and design. The system architecture is shown in figure 7, this system brings active learning of topics and strengthens the classification of the convolutional neural network model. The system was compared with SVM and convolutional neural network techniques individually to conclude and label new tweets. Nonetheless, the anticipating capacity of both approaches is defined due to irregular data sets.

Journal of Engineering, Electrical and Informatics
Vol.2, No.2 Juni 2022
e-ISSN: 2809-8706; p-ISSN: 2810-0557, Hal 23-37

Recommended Field Design Scheme

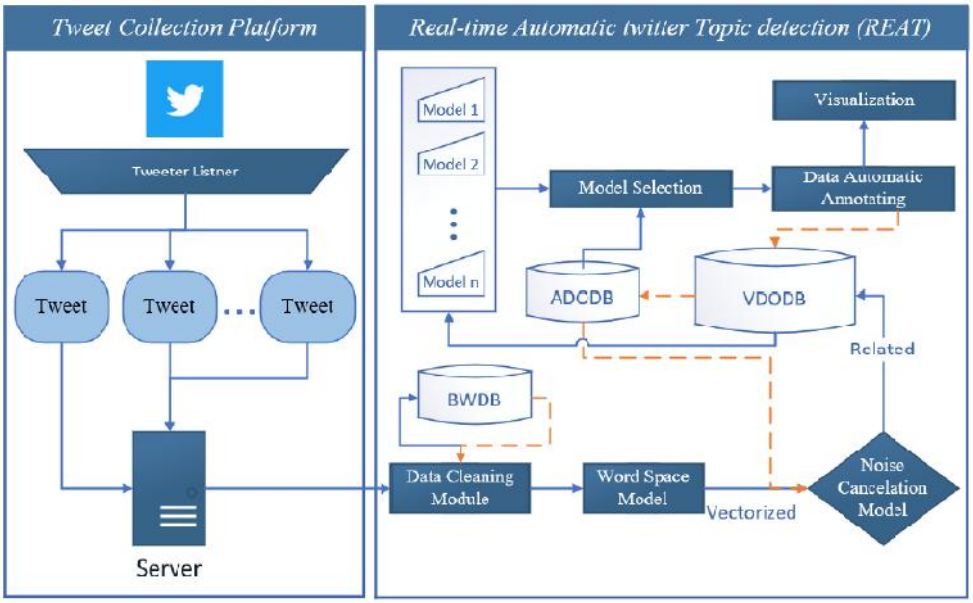

Figure 4. Real-Time Twitter Issue Framework Structure and Software Subjects

This system design has 2 important subjects, "Tweet Collection Platform" and "Real-time Automatic Twitter Topic detection" (REAT). The Tweet Collection Platform appliances listeners using the Tweepy API collect comments, and store them on the server for analysis by the “REAT” module. “REAT” decide 3 databases: “Bag of Word Data Base” (BWDB) to store tokens. “BWDB” will grow gradually, helping improve the tokenizer procedure, which is implemented in the Data cleaning module, and then "Vectorized DOcuments DataBase" (VDODB); this database will save the data after changing and removing noise, using "Word Space Model" and Noise elimination Model." And the last is Annotated Document DataBase” (ADODB) to increase the selector of topic modeling. This database stores annotated twitter. “ADODB” will help the framework calculate the homogeneity, completeness, and size of V. To generate ground truth and annotate random samples of data collected by assigning teams of characterized students and storing them in “ADODB”.

Noise elimination Model

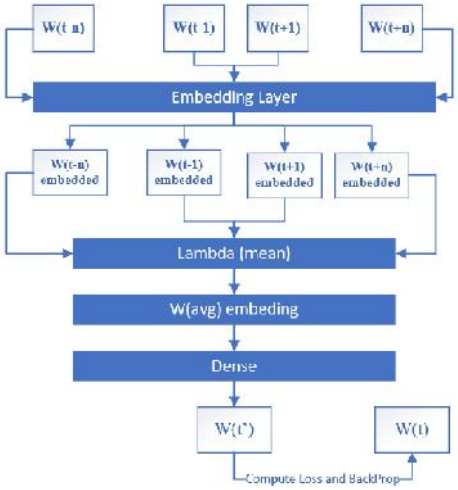

Figure 5. The layer of Deep Learning CB_OW

"The Continuous Bag of Words (CBOW) technique can clear up the problem of vocabulary size by crucial vector dimensions" (Mikolov et al., (2013)). The structure of the CB_OW deep learning model used as Word2Vec in this study serves on Figure 8. N serves as the window size, and W is the word, $W\ (\ t\ )$ will be the word in the document at position t , and creates a sequence $W\ (\ t{-}n\ )...\ W\ (\ t-1)\ W\ (\ t+1)\ W\ (\ t+n\ )$. V is the size of the Vocabulary and D is the dimension and makes the size of the matrix $V{*}D$. This matrix is known as the Enclose layer. To create the Enclose layer, a very important loss function needs to be defined. If the predicted target word is assumed to be $W\ (t')$ and the real word is $W\ (t)$, then this can decide a cost action. A heavy Layer (Vocabulary size) was applied with Softmax activation to predict the target $W\ (t')$. Table 3 shows the sample of input and a sample object by the window size, characterizing the two words before and after the target word. The end of restore weights on the heavy layer produces the word as a vector length (D).

Model Selection

At this stage, data is processed in different details, such as topic sentences, paragraphs, or articles. Latent Semantic Analysis was chosen as the basic algorithm, Latent Dirichlet Allocation as the most widely used technique, LDA_MAL-LET as an embellished adaptation of LDA, and Biterm Topic Modeling technique as a definite aim short message topic modeling, to calculate the scheme that proposed to choose the right approach for topic modeling.

Evaluation Metrics

Rosenberg & Hirschberg, (2007) "adopting V-Msd to calculate K-means performance through document grouping". They show that V-sizing can be done evidently on different data sets. Approaching this technique requires manual labeling. This study uses homogeneity, integrity, and V-Ms to calculate the techniques across topics. In this research, the metrics used are supervised to evaluate the resulting topic to pick a robust model.

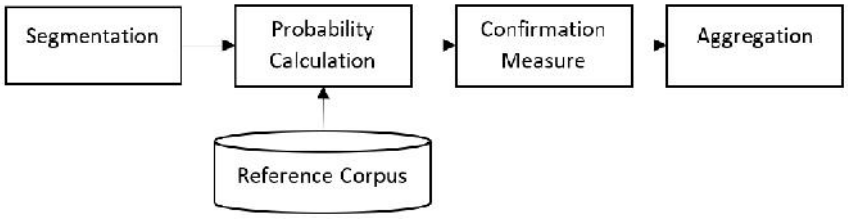

Figure 6. Consistency conclusion phase by distribution over the gathering

Analogy, Integrity, and V-size

Analogy describes the total of true labeled characters in one class. The goal is for each cluster to enclose members from one class. The harmonic average of these 2 pieces of evidence brings a V-Ms. Rosenberg & Hirschberg, (2007) described "this metric determines how close a cluster is to its ideal solution by examining the limited entropy of the class delivery given the expected clustering".

Journal of Engineering, Electrical and Informatics
Vol.2, No.2 Juni 2022
e-ISSN: 2809-8706; p-ISSN: 2810-0557, Hal 23-37

Description	Annotation
C as a Set of Classes	$C_{i_{1..n}}$
K as a Set of Cluster	$K_{j_{1..m}}$
Represent a member of class "c," which is an element of cluster "k"	a_{ck}
Number of Datapoint	N

Table 2. Glossary of Definitions in Homogeneity and Completeness Metrics

$$H(C|K) = -\sum_{k=1}^{|K|}\sum_{c=1}^{|C|}\frac{a_{ck}}{N}\log\frac{a_{ck}}{\sum_{c=1}^{|C|}a_{ck}}$$

$$H(C) = -\sum_{c=1}^{|C|}\frac{a_{ck}}{N}\log\frac{\sum_{k=1}^{|K|}a_{ck}}{n}$$

Based on this formula, homogeneity can be calculated using the formula:

$$P_{GL} = \begin{cases} if\ H(C|K) = 0 \quad Then\ h = 1 \\ else \qquad\qquad h = 1 - \dfrac{H(C|K)}{H(C)} \end{cases}$$

$$P_{GL} = \begin{cases} if\ H(K|C) = 0 \quad Then \quad c = 1 \\ else \qquad\qquad h = 1 - \dfrac{H(K|C)}{H(K)} \end{cases}$$

CONCLUSIONS AND FUTURE WORK

Conclusion

This research introduces a familiar structure for performing perfectly automated text compilation and sterilization, along with semi-automated case disclosure techniques, adopting a hybrid interpretation metric while assessing results regardless of the model chosen. The research product was measured by adopting V-Ms, agreement, integrity, and consistency to preferred several topics and models. In addition, trend-tracking samples demonstrating the rate of the investigation are anticipated by this groundwork. The hybrid model adopting assessment by the experts advances work in effective habitat. The continuous learning model also addresses the issue of shifting domains in date and area as in the analysis of the pandemic COVID-19, as a follow-up to this study. This study connects multiple data sources and creates integrated knowledge in healthcare, increasing the approach to the pharmaceutical ability for the public and experts. This study uses several approaches by the classification of semi-managed, deep learning, and transfer learning. The results of this study fulfill all the objectives defined in the establishment. This research makes a lot of improvements, as well as information collection, recommending 2 databases, creating a new commotion cancellation system for the social platform, developing a Deep Learning model for titled individual acceptance, and using transfer learning to conclude the model.

BIBLIOGRAPHY

Albishre, K., M. Albathan, and Y. Li, Effective 20 newsgroups dataset cleaning. The 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015. 3: p. 98--101.

Bates, DW, et al., Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Affairs, 2014. 33(7): p. 1123-1131.

Blei, DM, AY Ng, and MI Jordan, Latent Dirichlet allocation. Journal of machine Learning research, 2003. 3(1): p. 993--1022.

Bodenreider, O., The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 2004. 32(90001): p. 267D-270.

Campos, D., S. Matos, and JL Oliveira, Gimli: open source and high-performance biomedical name recognition. BMC Bioinformatics, 2013. 14(1): p. 54.

Cawley, GC, and NL Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research, 2010. 11: p. 2079-2107.

Chapman, BE, et al., Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. Journal of Biomedical Informatics, 2011. 44(5): p. 728-737.

Chapman, W., J. Dowling, and D. Chu. ConText: An algorithm for identifying contextual features from clinical text. in Biological, translational, and clinical language processing. 2007.

Chiu, JPC and E. Nichols, Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 2016. 4: p. 357--370.

Choi, E., et al., Medical concept representation of learning from electronic health records and its application on heart failure prediction. arXiv preprint arXiv:1602.03686, 2016.

Deerwester, S., et al., Indexing by latent semantic analysis. Journal of the American society for information science, 1990. 41(6): p. 391--407.

DeHart, K. and J. Holbrook, Emergency department applications of digital dictation and natural language processing. The Journal of ambulatory care management, 1992. 15(4): p. 18-23.

Doan, RI, R. Leaman, and Z. Lu, NCBI disease corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics, 2014. 47: p. 1--10.

Dumais, ST, Latent semantic analysis. Annual review of information science and technology, 2004. 38(1): p. 188--230.

Gerner, M., G. Nenadic, and CM Bergman, LINNAEUS: A species name identification system for biomedical literature. BMC Bioinformatics, 2010. 11(1): p. 85.

Guo, L., et al., Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. Journalism \& Mass Communication Quarterly, 2016. 93(2): p. 332--359.

Huang, Z., W. Xu, and K. Yu, Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991, 2015.

Hwang, M.-H., et al., Spatiotemporal transformation of social platform geostreams: a case study of Twitter for flu risk analysis. The 4th ACM SIGSPATIAL International Workshop on GeoStreaming, 2013: p. 12--21.

Kelly, L., et al. Overview of the share/clef eHealth evaluation lab 2014. in International Conference of the Cross-Language Evaluation Forum for European Languages. 2014. Springer.

Kim, J.-D., et al., Introduction to the bio-entity recognition task at JNLPBA. 2004 :p. 70--75.

Kingma, DP and J. Ba, Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. in Ijcai. 1995. Montreal, Canada.

Krestel, R., P. Fankhauser, and W. Nejdl, Latent Dirichlet allocation for tag recommendation. The third ACM conference on Recommender systems, 2009: p. 61--68.

Krstajic, D., et al., Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of cheminformatics, 2014. 6(1): p. 1-15.

Kuperman, GJ, et al., Medication-related clinical decision support in computerized provider order entry systems: a review. Journal of the American Medical Informatics Association, 2007. 14(1): p. 29-40.

Lafferty, J., A. McCallum, and FCN Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

Laylavi, F., A. Rajabifard, and M. Kalantari, Event relatedness assessment of Twitter messages for emergency response. Information processing \& management, 2017. 53(1): p. 266--280.

Leaman, R. and G. Gonzalez, BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput, 2008: p. 652-63.

Li, J., et al., BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Databases, 2016. 2016.

Liu, F., C. Weng, and H. Yu, Natural language processing, electronic health records, and clinical research, in Clinical Research Informatics. 2012, Springer. p.s. 293-310.

Ma, X. and E. Hovy, End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354, 2016.

Marasovi, Srl4orl: Improving opinion role labeling using multi-task learning with semantic role labeling. arXiv preprint arXiv:1711.00768, 2017.

McCallum, A., D. Freitag, and FCN Pereira, Maximum Entropy Markov Models for Information Extraction and Segmentation. 2000.17 :p. 591--598.

McCallum, AK, Mallet: A machine learning for language toolkit. http://mallet. cs. umass. edu, 2002.

Mi, H. and P. Thomas, PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. Methods Mol Biol, 2009. 563: p. 123-40.

Mikolov, T., et al., Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

Mikolov, T., et al., Recurrent neural network based language model. 2010.

Mowery, D., Developing a Clinical Linguistic Framework for Problem List Generation from Clinical Text. 2014, University of Pittsburgh.

Passos, A., V. Kumar, and A. McCallum, Lexicon infused phrase enclosed for named entity resolution. arXiv preprint arXiv:1404.5367, 2014.

Pennington, J., R. Socher, and CD Manning, Glove: Global vectors for word representation. 2014 :p 1532--1543.

Pennington, J., R. Socher, and CD Manning. Glove: Global vectors for word representation. in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

Prier, KW, et al. Identifying health-related topics on twitter. in International conference on social computing, behavioral-cultural modeling, and prediction. 2011. Springer.

Rathore, MM, et al., Advanced computing model for geosocial platform using big data analytics. Multimedia Tools and Applications, 2017. 76(23): p. 24767--24787.

Ratinov, L. and D. Roth, Design challenges and misconceptions in named entity recognition. 2009:p.p. 147--155.

Ratnaparkhi, A., A maximum entropy model for part-of-speech tagging. 1996.

Rder, M., A. Both, and A. Hinneburg, Exploring the space of topic coherence measures. The eighth ACM international conference on Web search and data mining, 2015: p. 399--408.

Reddy, CK, and CC Aggarwal, Healthcare data analytics. Vol. 36. 2015: CRC Press.

Rosenberg, A. and J. Hirschberg, V-Ms: A conditional entropy-based external cluster evaluation measure. The 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), 2007: p. 410--420.

Santos, CD and B. Zadrozny, Learning character-level representations for part-of-speech tagging. 2014 :p 1818--1826.

Scanfeld, D., V. Scanfeld, and EL Larson, Dissemination of health information through social networks: Twitter and antibiotics. American journal of infection control, 2010. 38(3): p. 182-188.

Smith, L., et al., Overview of BioCreative II gene mention recognition. Genome Biology, 2008. 9(S2): p. S2.

Suominen, H., et al. Overview of the ShARe/CLEF eHealth evaluation lab 2013. in International Conference of the Cross-Language Evaluation Forum for European Languages. 2013. Springer.

Surian, D., et al., Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. Journal of medical Internet research, 2016. 18(8): p. e232.

Svenstrup, D., JM Hansen, and O. Winther, Hash enclosed for efficient word representations. arXiv preprint arXiv:1709.03933, 2017.

Terry, M., Twittering Healthcare: Social platform and Medicine. Telemedicine and e-Health, 2009. 15(6): p. 507-510.

Thom, D., et al., Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages. The 2012 IEEE Pacific Visualization Symposium, 2012: p. 41--48.

Tran, T., et al., Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). Journal of biomedical informatics, 2015. 54: p. 96-105.

Yan, X., et al., A bitterm topic model for short texts. The 22nd international conference on the World Wide Web, 2013: p. 1445--1456.

Yen, S.-J., et al., A support vector machine-based context-ranking model for question answering. Information Sciences, 2013. 224: p. 77--87.

Zaremba, W., I. Sutskever, and O. Vinyals, Recurrent neural network regularization. arXiv preprint arXiv:1409.2329, 2014.

Zhao, L., et al., Spatiotemporal event forecasting in social platform. the 2015 SIAM international conference on data mining, 2015: p. 963--971.

Zhou, P., et al., Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. arXiv preprint arXiv:1611.06639, 2016.