International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

DEEP LEARNING APPROACH TO IMPLEMENT PLAGIARISM CHECKER

1. Joshi Padma N, Associate Professor, Department of CSE, Sreyas Institute of Engineering and Technology,Telangana, India, padmajoshi@sreyas.ac.in

2. B. Swathi, Department of CSE, Sreyas Institute of Engineering and Technology, Telangana, India, bonthaswathi2002@gmail.com

3. A. Manisha, Department of CSE, Sreyas Institute of Engineering and Technology, Telangana, India. manishareddyanagu@gmail.com

4. A. Sarika, Department of CSE, Sreyas Institute of Engineering and Technology, Telangana, India, sarikachinnu2407@gmail.com

5. K. Rahul Yadav, Department of CSE, Sreyas Institute of Engineering and Technology, Telangana, India, rahulyadavkanyamoni@gmail.com

ABSTRACT: Natural Language Sentence Matching (NLSM) is one of the important and challenging tasks in Natural Language Processing where the task is to identify if a sentence is a paraphrase of another sentence in a given pair of sentences. Paraphrase of a sentence conveys the same meaning but its structure and the sequence of words varies. It is a challenging task as it is difficult to infer the proper context about a sentence given its short length. Also, coming up with similarity metrics for the inferred context of a pair of sentences is not straightforward as well. Whereas, its applications are numerous. This work explores various machine learning algorithms to model the task and also applies different input encoding scheme. Specifically, we created the models using Logistic Regression, Support Vector Machines, and different architectures of Neural Networks. Among the compared models, as expected, Recurrent Neural Network (RNN) is best suited for our paraphrase identification task. Also, we propose that Plagiarism detection is one of the areas where Paraphrase Identification can be effectively implemented.

Keywords *– NLP, NLSM, RNN, LSTM, Plagiarism*

1. INTRODUCTION

Paraphrase identification is the task of identifying if a sentence is a paraphrase of another one. It is one of the challenging tasks in Natural Language Processing. It requires representing a text in some form taking its context into consideration and formulating a metric to express the similarity between a pair of texts. The given pair of sentences or texts may look almost similar in terms of its syntactical structure but a presence of a single word or phrase may convey entirely different or opposite meanings. On the other hand, there are various applications of paraphrase identification. One of them can be automatically removing the duplicate questions in online QA forums like Quora. There are different versions of the same question in such online question-answer forums conveying the same meaning. Traditional coding would require creating billions of conditions to accurately assess whether or not two sentences are semantically the same. Another application can be the plagiarism detection task. Current applications for plagiarism essentially just check the syntax. With a quick web search, one can find many waysto circumvent plagiarism detection by switching out select words or using an article rewriter. Paraphrase identification is suggested as an application-independent framework for measuring semantic equivalence. In terms of identifying

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

duplicate questions, according to [1], it explains that if two questions can be concluded with the same answer, both questions are semantically equivalent. The identification of semantically equivalent sentences has many applications in natural language understanding which ranges from paraphrase recognition to evaluating machine translation. There are a few challenges when it comes to processing the texts using machines. First, the computer finds it hard to recognize different words and their meanings. For example, when speaking of the companies, “Microsoft” or “Apple”, the computer might mistake “Apple” as a fruit instead of a company. This problem occurs because generally, machines fail to figure out the context depicted in the text. In a given natural language sentence, there are various relationships among the words. So capturing this relationship is essential to completely understand the semantic of that sentence.

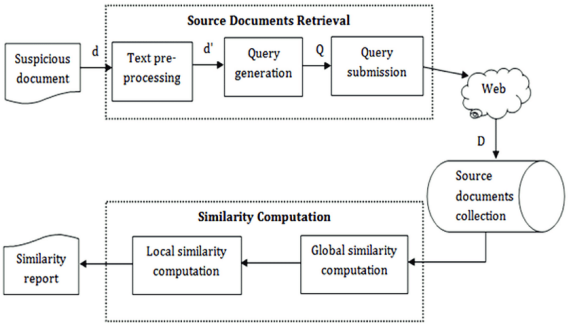

Fig.1: Example figure

The word "plagiarism" has originated from the word "plagiarius" which means "kidnapper" in Latin. It is tough to give an exact definition for the plagiarism, but according to the Oxford dictionary simple meaning of the plagiarism is "The practice of taking someone else's work or ideas and passing them off as one's own". The main reason behind the vast increase in plagiarism is due to the open and unlimited availability of resources over the Internet. Plagiarism is a theft of intellectual property, and it has been in practice in one way or another since the human has produced work of art and research [1]. There are four broader categories of plagiarism and can be listed as follows [2] [3]: 1) Accidental: due to the lack of understanding about plagiarism and its consequences. 2) Unintentional: the availability of abundant material influences one's thoughts and the same ideas may be expressed in similar verbal or written expressions. 3) Intentional: a deliberate act of copying complete or part of someone else's work without giving proper credit to the original creator. 4) Self-plagiarism: copying from self-published work without referring to the original one.

2. LITERATURE REVIEW

Detecting Duplicate Questions with Deep Learning

In this paper, we explore methods of determining semantic equivalence between pairs of questions using a dataset released by Quora. Our deep learning approach to this problem uses a Siamese GRU neural network to encode each sentence, and we experiment with a variety of distance measures to predict equivalence based on the sentence vector outputs of the neural network. We find that while logistic regression on the pure distance measures produces decent results, feeding a concatenation of different transformations of the output sentence vectors through another set of neural network layers yields significantly improves performance to a level comparable to current state-of-the-art models. In addition, we demonstrate data augmentation techniques that can be used to improve Siamese neural network model performance.

Using deep learning for short text understanding

Classifying short texts to one category or clustering semantically related texts is challenging, and the importance of both is growing due to the rise of microblogging platforms, digital news feeds, and the like. We can accomplish this classifying and clustering with the help of a deep neural network which produces compact binary representations of a short text, and can assign the same category to texts that have similar binary representations. But problems arise when there is little contextual information on the short texts, which makes it difficult for the deep neural network to produce similar binary codes for semantically related texts.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

We propose to address this issue using semantic enrichment. This is accomplished by taking the nouns, and verbs used in the short texts and generating the concepts and co-occurring words with the help of those terms. The nouns are used to generate concepts within the given short text, whereas the verbs are used to prune the ambiguous context (if any) present in the text. The enriched text then goes through a deep neural network to produce a prediction label for that short text representing it's category.

Gated Self-Matching Networks for Reading Comprehension and Question Answering

In this paper, we present the gated selfmatching networks for reading comprehension style question answering, which aims to answer questions from a given passage. We first match the question and passage with gated attention-based recurrent networks to obtain the question-aware passage representation. Then we propose a self-matching attention mechanism to refine the representation by matching the passage against itself, which effectively encodes information from the whole passage. We finally employ the pointer networks to locate the positions of answers from the passages. We conduct extensive experiments on the SQuAD dataset. The single model achieves 71.3% on the evaluation metrics of exact match on the hidden test set, while the ensemble model further boosts the results to 75.9%. At the time of submission of the paper, our model holds the first place on the SQuAD leaderboard for both single and ensemble model.

Acquiring Predicate Paraphrases from News Tweets

We present a simple method for evergrowing extraction of predicate paraphrases from news headlines in Twitter. Analysis of the output of ten weeks of collection shows that the accuracy of paraphrases with different support levels is estimated between 60-86%. We also demonstrate that our resource is to a large extent complementary to existing resources, providing many novel paraphrases. Our resource is publicly available, continuously expanding based on daily news

A Semantic Similarity Approach to Paraphrase Detection

This paper presents a novel approach to the problem of paraphrase identification. Al-though paraphrases often make use of syn-onymous or near synonymous terms, many previous approaches have either ignored or made limited use of information about simi-larities between word meanings. We present an algorithm for paraphrase identification which makes extensive use of word similar-ity information derived from WordNet (Fell-baum, 1998). The approach is evaluated us-ing the Microsoft Research Paraphrase Cor-pus (Dolan et al., 2004), a standard resource for this task, and found to outperform previ-ously published methods.

A Semantic Approach to IE Pattern Induction

This paper presents a novel algorithm for the acquisition of Information Extraction patterns. The approach makes the assump- tion that useful patterns will have simi- lar meanings to those already identified as relevant. Patterns are compared using a variation of the standard vector space model in which information from an on- tology is used to capture semantic sim- ilarity. Evaluation shows this algorithm performs well when compared with a previously reported document-centric ap- proach.

3. METHODOLOGY

The existing plagiarism detection methods work on identical text matching strategies, without considering the core of the knowledge or how this knowledge is developed. They are inherently limited by their rigid assumptions that plagiarists literally copy and paste whole sentences or paragraphs of text from other authors directly into their documents. Most of the existing plagiarism detection methods are limited and have a number of shortcomings in detecting many types of plagiarism cases (e.g., syntactical or semantics changes).

Limitations:

The ease of sharing online information in this age of digital communication has encouraged the misuse of

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

text and the prevalence of plagiarism. Academic bodies and scientific publishing companies are playing an active role in detecting plagiarism in order to maintain the integrity of academic publications ML Techniques are used to find out the plagiarized content but it cannot find the similarities between small sentences and works only on big sentences. The techniques which are used cannot able to find the semantic and syntactical changes.

We try to perform paraphrase identification using various machine learning models and make a performance comparison among these models. In Recent years, Recurrent Neural Networks (RNNs) have proven to be very successful in machine learning tasks that relate to Natural Language. As Natural Language can be represented as a sequence of tokens(characters, words or phrases), and in general the preceding tokens affect the occurrence of next token in the sequence, RNNs, which work by taking the feedback of previous time-steps output to generate the output for subsequent time-steps, works very well. Specifically, we devise a Logistic Regression model which is the simplest machine learning model for classification task and then go on to implement a relatively more complex model: Support Vector Machines. Lastly, we will develop various models using Neural Networks including RNN model.

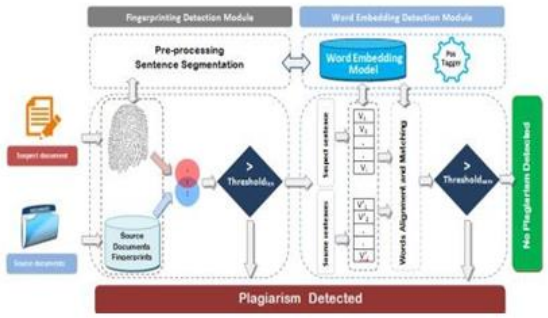

Fig.2: System architecture

MODULES:

To implement aforementioned project we have designed following modules

- Data exploration: using this module we will load data into system
- Processing: Using the module we will read data for processing
- Splitting data into train & test: using this module data will be divided into train & test
- Model generation: Build model – logistic regression, SVM, LSTM and neural network and calculate accuracy values.
- User input: Using this module will give input for prediction
- Prediction: final predicted displayed

4. IMPLEMENTATION

Logistic Regression:

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

the classification. The below image is showing the logistic function:

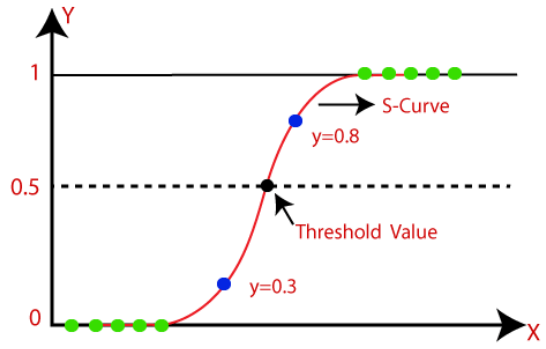

Fig.3: Logistic Regression

Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

We know the equation of the straight line can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}\text{; 0 for y= 0, and infinity for y=1}$$

But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

The above equation is the final equation for Logistic Regression.

Support Vector Machine:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. 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:

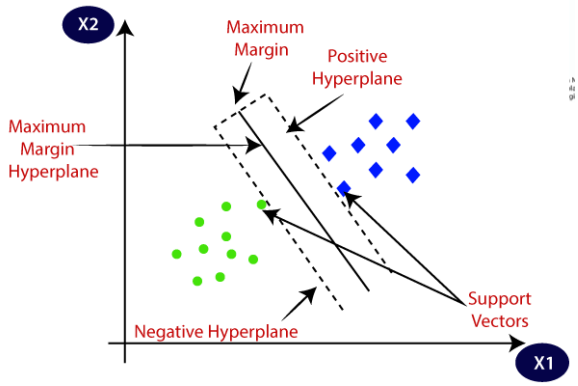

Fig.4: SVM

SVM can be of two types:

Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

Non-linear SVM: Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

LSTM:

LSTM is a type of Neural Network used in the field of Deep Learning. LSTM stands for Long-Short-term-Memory. LSTM is an improved version of the RNN(Recurrent Neural Network). LSTM is mainly used in Time series and Sequence data because RNN doesn't perform efficiently as the gap length rises. LSTM differs from conventional Feedforward Networks as it uses previous data and its output to affect the current predictions. LSTM is also better at retaining information for longer periods when compared with RNN. Long Short Term Memory uses Gated Cells to remember or forget previous information.

LSTM is a cell that consists of 3 gates. A forget gate, input gate, and output gate. The gates decide which information is important and which information can be forgotten. The cell has two states Cell State and Hidden State. They are continuously updated and carry the information from the previous to the current time steps. The cell state is the “long-term” memory, while the hidden state is the “short-term” memory.

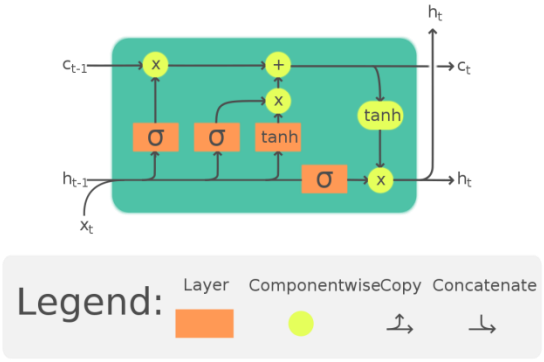

Fig.5: LSTM

Neural Networks:

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. Neural networks can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria. The concept of neural networks, which has its roots in artificial intelligence, is swiftly gaining popularity in the development of trading systems.

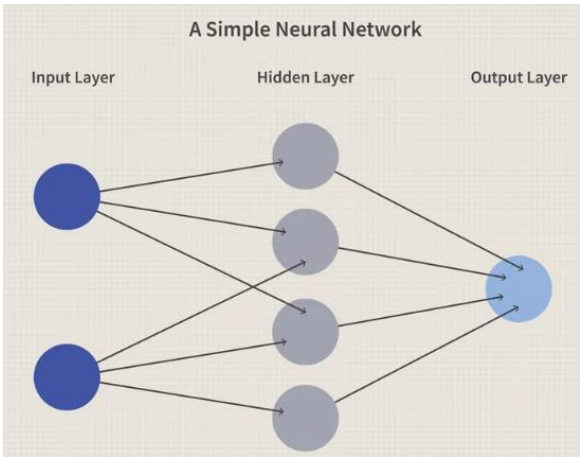

Fig.6: Neural Networks

5. EXPERIMENTAL RESULTS

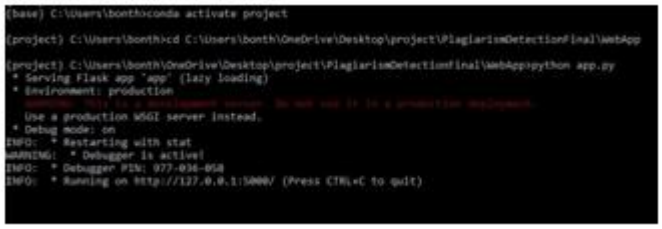

Fig.7: Command Prompt

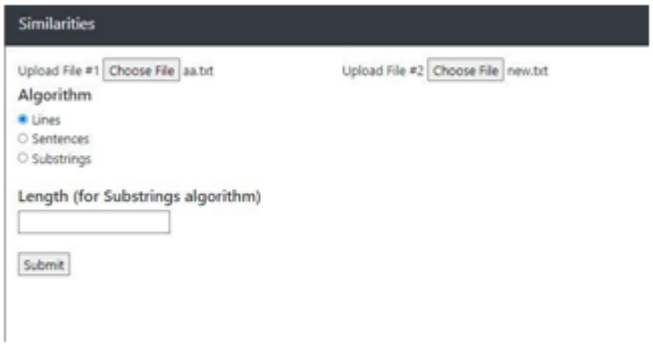

Fig.8: Web page

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Fig.9: Comparison Between Lines

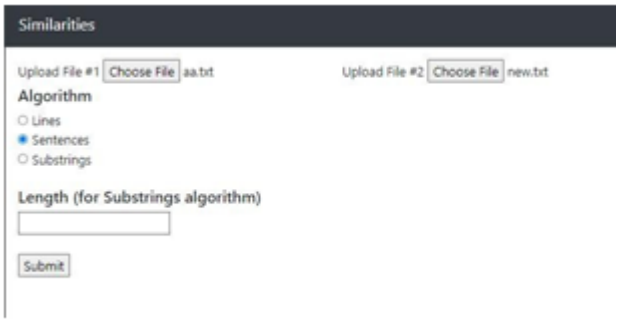

Fig.10: Choosing algorithm

Fig.11: Comparison between sentences

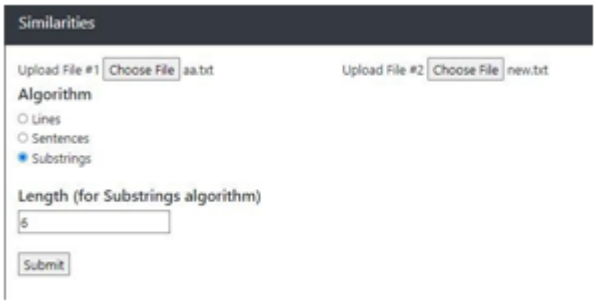

Fig.12: Choosing algorithm

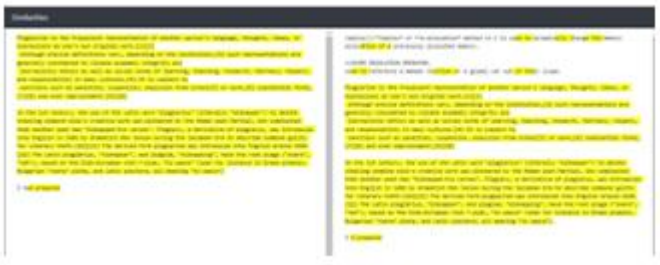

Fig.13: Comparison between substrings

6. CONCLUSION

Paraphrase identification can be used in many applications. One of them, we propose in Plagiarism detection. One of the main problems with plagiarism checkers is they check for syntactical structures only instead of semantic meaning. Our model can be used to develop a plagiarism detection system where a simple rewrite of a text will be flagged as plagiarized. We have developed a simple application using our model for this purpose. In conclusion, Various machine learning algorithms were implemented for paraphrase identification tasks. Specifically, we used Logistic Regression, Support Vector Machine, and Neural Networks. As expected, recurrent neural networks (RNN) were found to produce the most accurate results. Furthermore, we propose that Paraphrase Identification can be implemented for plagiarism detection effectively and also developed a simple application for the demonstration purpose.

7. FUTURE WORK

Artificial intelligence is in use in many facets of our lives, from our smartphones and smart digital assistants to our smart home devices. Apart from those objects. Applications such as these utilize this advanced technology to detect content that may bear any similarity to previously published texts. No matter how they have been rearranged or paraphrase carefully, a powerful AI system can determine whether or not content has been copied from another source or has duplicates in the web across different platforms.

There will come a day when an Artificial General Intelligence (AGI) waxes eloquent, making it difficult to discern if the AGI is plagiarizing. For now, the fundamental orientation of current Artificial Intelligence (AI) is to detect meaning and patterns, as well as to draw inferences, from the data the AI is presented. The amount of original prose AIs produce to communicate findings is very limited. They do uncover otherwise obscure relationships, but how they do it is construed as 'original thought' without the AI 'knowing' about parallel works.

REFERENCES

[1] Y. Homma, S. Sy, and C. Yeh, "Detecting duplicate questions with deep learning," in Proceedings of the International Conference on Neural Information Processing Systems (NIPS, 2016.

[2] J. Zhan and B. Dahal, "Using deep learning for short text understanding," Journal of Big Data, vol. 4, no. 1, p. 34, 2017.

[3] W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou, "Gated selfmatching networks for reading comprehension and question answering," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 189– 198.

[4] V. Shwartz, G. Stanovsky, and I. Dagan, "Acquiring predicate paraphrases from news tweets," in Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017), 2017, pp. 155–160.

[5] S. Fernando and M. Stevenson, "A semantic similarity approach to paraphrase detection," in Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics, 2008, pp. 45–52.

[6] M. Stevenson and M. A. Greenwood, "A semantic approach to ie pattern induction," in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005, pp. 379–386.

[7] R. Yangarber, "Counter-training in discovery of semantic patterns," in Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003, pp. 343–350.

[8] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in Proceedings of the 3rd international workshop on Search and mining user-generated contents. ACM, 2011, pp. 37–44.

[9] A. Rajkumar and A. Chitra, "Paraphrase recognition using neural network classification," International Journal of Computer Applications, vol. 1, no. 29, pp. 42–47, 2010.

[10] V. Hatzlvassiloglou, J. L. Klavans, and E. Eskin, "Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning," in 1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora, 1999.

[11] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," arXiv preprint arXiv:1606.05250, 2016.

[12] C. Joao, D. Gael, and B. Pavel, "New functions for unsupervised ¨ asymmetrical paraphrase detection," Journal of Software, vol. 2, no. 4, pp. 12–23, 2007.

[13] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Semantic enrichment of twitter posts for user profile construction on the social web," in Extended semantic web conference. Springer, 2011, pp. 375–389.

[14] H. He, K. Gimpel, and J. Lin, "Multi-perspective sentence similarity modeling with convolutional neural networks," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1576–1586.

[15] N. R. R. M. M. S. M. B. J. Z. L. G. P. O. Felix Zhan, Anthony Martinez, "Beyond cumulative

sum charting in non-stationarity detection and estimation," IEEE Access, 2019.

[16] Z. J. Schwob, M. and D. A., "Modeling cell communication with time-dependent signaling hypergraphs," IEEE/ACM Transactions on Computational Biology and Bioinformatics, p. doi: 10.1109/TCBB.2019.2937033, 2019.

[17] C. Chiu and J. Zhan, "Deep learning for link prediction in dynamic networks using weak estimators," IEEE Access, vol. 6, no. 1, pp. 35 937 – 35 945, 2018.

[18] M. Bhaduri and J. Zhan, "Using empirical recurrences rates ratio for time series data similarity," IEEE Access, vol. 6, no. 1, pp. 30 855– 30 864, 2018.

[19] J. Wu, J. Zhan, and S. Chobe, "Mining association rules for low frequency itemsets," PLOS ONE, vol. 13, no. 7, 2018.