IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MULTI MODEL NATURAL LANGUAGE PROCESSING

Archit Shukla[1] Ganesh Makkina[2] Gaurav Kumar[3] Gaurav Kumar[4]

***Abstract*— Recent advances in Big Data has prompted health care practitioners to utilize the data available on social media to discern sentiment and emotions expression. Health Informatics and Clinical Analytics depend heavily on information gathered from diverse sources. Traditionally, a healthcare practitioner will ask a patient to fill out a questionnaire that will form the basis of diagnosing the medical condition. However, medical practitioners have access to many sources of data including the patients writings on various media. Natural Language Processing (NLP) allows researchers to gather such data and analyze it to glean the underlying meaning of such writings. The field of sentiment analysis (applied to many other domains) depend heavily on techniques utilized by NLP. This work will look into various prevalent theories underlying the NLP field and how they can be leveraged to gather users sentiments on social media. Such sentiments can be culled over a period of time thus minimizing the errors introduced by data input and other stressors. Furthermore, we look at some applications of sentiment analysis and application of NLP to mental health. The reader will also learn about the NLTK toolkit that implements various NLP theories and how they can make the data scavenging process a lot easier.**

I. INTRODUCTION

Computing the similarity between two text documents is a common task in NLP, with several practical applications. It has commonly been used to, for example, rank results in a search engine or recommend similar content to readers. Since text similarity is a loosely-defined term, we'll first have to define it for the scope of this article. After that, we'll explore two different ways of computing similarity and their pros and cons.

II. VARIOUS TECHNIQUES

Our first step is to define what we mean by similarity. We'll do this by starting with two examples. Let's consider the sentences:

- The teacher gave his speech to an empty room
- There was almost nobody when the professor was talking

Although they convey a very similar meaning, they are written in a completely different way. In fact, the two sentences just have one word in common ("the"), and not a really significant one at that. Nevertheless, it's safe to say that we'd want an ideal similarity algorithm to return a high score for this pair.

Now let's change the sentences a little bit:

- The teacher gave his speech to an empty a full room
- There was almost nobody when the professor was talking We only changed two words, yet the two sentences now have an opposite meaning.

A. *COSINE SIMILARITY*

Cosine similarity measures the similarity be- tween two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis. Note that the cosine similarity measure equates the lack of ratings as zero value ratings, which can also be considered as a low rating. This assumption works fine for the applications for where the user has purchased the item or not. In the movie recommendation case, this assumption can yield the wrong results because the lack of rating does not mean that the user dislikes the movie. Some users are more generous in giving ratings than others who are more critical. The user's bias in giving ratings skews the similarity score between users.

B. TF-IDF (Term Frequency - Inverse Document Fre- quency)

TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a col- lection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents. TF-IDF (term frequency-inverse document frequency) was invented for document search and information retrieval. So, words that are common in every document, such as this, what, and if, rank low even though they may appear many times, since they don't mean much to that document in particular.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

FIG 1:(a) To put it in more formal mathematical terms, the TF-IDF score for the word t in the document d from the document set D is calculated as shown above

C. PREPROCESSING TECHNIQUES

Preprocessing is an important phase use to manipulate real world data into an un- derstandable format. Surely, the real-world data have been often incomplete, noisy in specific behavior. In other words, most of data that we wish to analyze from real world by using data mining techniques are incomplete and inconsistent (containing errors, outlier values). Hence, the preprocessing methods are required before ap- plying data mining techniques to improve the quality of the data, thus assisting to enhance the accuracy and efficiency of resulting data mining task. The preprocess- ing techniques are vital and important in network traffic analysis due to the patterns of network traffic which have different type of format and dimensionality.

- Discretization method
- Feature Selection Method

D. NEURAL NETWORK

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.

E. Neural Networks in NLP Applications

Transfer learning is aimed to make use of valu- able knowledge in a source domain to help model performance in a target domain. It is particularly important to neural net- works, which are very likely to be overfitting. In some fields like image processing, many studies have shown the effec- tiveness of neural network-based transfer learn- ing. For neural NLP, however, existing studies have only casually applied transfer learning, and conclusions are inconsistent. In this paper, we conduct systematic case studies and provide an illuminating picture on the transferability of neural networks in NLP.

F. Word Embedding

Word embedding is one of the most popular representation of document vocabulary. It is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc.

G. Word2Vec

Word2Vec is a more recent model that em- beds words in a lower-dimensional vector space using a shallow neural network. The result is a set of word-vectors where vectors close together in vector space have similar meanings based on context, and word-vectors distant to each other have differing meanings. For example, strong and powerful would be close together and strong and Paris would be relatively far.

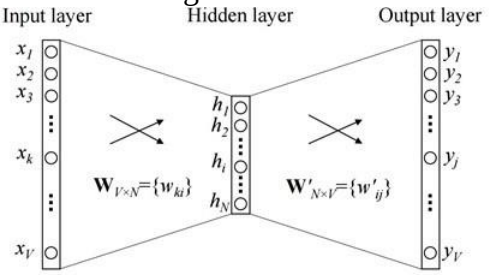


Figure 1: A simple CBOW model with only one word in the context

III. WORKING

A. *Word Embeddings use case scenarios*

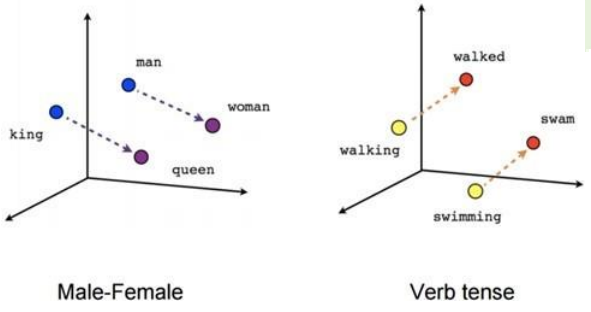
Since word embeddings or word Vectors are numerical representations of contextual similarities between words, they can be manipulated and made to perform amazing tasks

B. *Semantic Analysis*

Semantic analysis describes the process of understanding natural language–the way that humans communicate–based on meaning and context. The semantic analy- sis of natural language content starts by reading all of the words in content to capture the real meaning of any text. It identifies the text elements and assigns them to their logical and grammatical role. It analyzes context in the surrounding text and it ana- lyzes the text structure to accurately disambiguate the proper meaning of words that have more than one definition.

C. *Grammatical error detection*

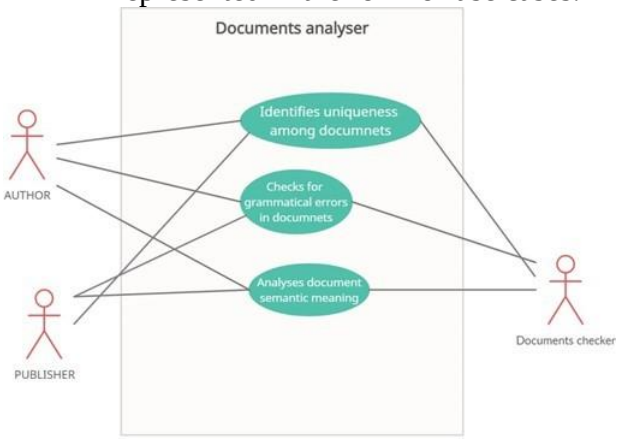
To retrieve writing suggestions for detected errors, we design queries for each edit type to search for more plausible corrections using Linggle, a linguistic search en- gine on a web- based dataset of one trillion words (Boisson et al., 2013). Linggle has different query functions and operators to search word usage in context.



Male-Female Verb tense

IV. USE CASE DIAGRAM

A use case diagram at its simplest is a representation of a user's interaction with the sys- tem that shows the relationship between the user and the different use cases in which the user is involved. In our project the primary user of the system is the Customer. The diagram is used to tell the actors reaction to various sce- narios represented in the form of use cases.



V. CONCLUSION

Structural data obtained using the unstructured model of the data using methods of text and web mining be brought, and from there obtained structural models were compared. The results obtained, text and web mining methods using the obtained model is more successful shows. Unstructured data model to integrate the quality of information can be removed and this result also shows that the unexpected results. Potentially used in the world 80types of data considered, the use of this data will certainly add value to research. The other impor- tant issue is evaluation unstructured data analysis in text mining methods and NLP techniques. Text mining try to finding patterns in textual unstruc- tured files based on contents. NLP try to reach concepts of texts via specific algorithms.

REFERENCES

[1] 1. Y.H. Wang, C.H. Lin, "An English sentence parser for grammar error detection", 2002 IEEE Region 10 Conference on Computers, Communica- tions, Control and Power En- gineering. TENCOM '02. Proceedings, DOI: 10.1109/TEN-CON.2002.1181309

[2] [2] Oguzhan Menemencioglu,Ilhami M. Orak, "A Review on Semantic Web and Recent Trends in Its Applications", 2014 IEEE International Conference on Semantic Computing, DOI: 10.1109/ICSC.2014.57

[3] [3] Bo-Hao Su, Ta-Wen Kuan, Shih-Pang Tseng, "Im- proved TF-IDF weight method based on sentence sim- ilarity for spoken dialogue system", 2016 Inter- na- tional Conference on Orange Technologies (ICOT), DOI: 10.1109/ICOT.2016.8278973

[4] [4]Deep Neural Networks for No-Reference and Full-Reference Image Qual- ity Assessment, DOI: 10.1109/TIP.2017.2760518

[5] [5] Resource-based Natural Language Processing, DOI: 10.1109/NLPKE.2007.4368002

[6] [6] Hitoshi Isahara, "Resource-based Natural Language Processing", 2007 In- ternational Conference on Natural Language Processing and Knowledge En- gineering DOI: 10.1109/NLPKE.2007.4368002

[7] [7] Sebastian Bosse, Dominique Maniry, "Deep Neu- ral Networks for No-Reference and Full-Reference Im- age Quality Assessment", IEEE Transactions on Image Processing (Volume: 27, Issue: 1, Jan. 2018), DOI: 10.1109/TIP.2017.2760518

[8] [8] O. Abdel-hamid, L. Deng and D. Yu, Exploring Con- volutional Neural Net- work Structures and Optimization Techniques for Speech Recognition, pp. 3366-3370, August 2013.

[9] [9] A. Krizhevsky, I. Sutskever and G.E. Hinton, "Imagenet classification with deep convolutional neural networks", Ad- vances in neural information process- ing systems, pp. 1097- 1105, 2012.

[10] [10] Y. Taigman, M. Yang, M.A. Ranzato and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701- 1708, 2014.

[11] [11] Damar Zaky, Ade Romadhony, "An LSTM-based Spell Checker for In- donesian Text", 2019 International Confer- ence of Advanced Informatics: Con- cepts, Theory and Appli- cations (ICAICTA), DOI: 10.1109/ICAICTA.2019.8904218