Artificial Intelligence based Semantic Text Similarity for RAP Lyrics

Chandra J
Department of Computer Science
CHRIST (Deemed to be University)
Banglore, India
chandra.j@christuniversity.in

Akshay Santhanam
Department of Computer Science
CHRIST (Deemed to be University)
Banglore, India
santhanamakshay22@gmail.com

Alwin Joseph
Department of Computer Science
CHRIST (Deemed to be University)
Banglore, India
alwin.joseph@mca.christuniversity.in

2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) 978-1-7281-4142-8/20/$31.00 ©2020 IEEE 10.1109/ic-ETITE47903.2020.211

Abstract— **Data mining is the primary method of gathering large volumes of knowledge. To make use of such data to implementation requires the use of effective machine learning strategies. Semantic Textual Similarity is one of the primary machine learning strategies. At its core, semantic textual similarity is the identification of words with similar context. Initial work in STS involved text reuse, word search among others. The proposed research work uses a specific method of determining textual similarity using Google's Word2Vec framework and the Continuous-bag-of-words algorithm for identifying word similarity in rap records. A large data set consisting of over 50,000 rap records is used. The key aspect of proposed methodology is to determine the words with similar context and cluster them into different word clusters also called bags. To achieve the desired result, the dataset is first processed to obtain the features. Once the features are selected, a model is generated by passing the data onto the Word2Vec framework. The research work on semantic textual similarity was repeated across three different runs, with the data set size changing in every run. At the end of each the accuracy of similarity obtained by the model was determined. The current research work has achieved average accuracy as 85%.**

Keywords— Word2Vec, Semantic Textual Similarity (STS), natural language processing, bag-of-words, Continuous Bag of Words

I. INTRODUCTION

Rap is a musical genre famous for its fast paced, hard hitting, beat based sounds and rhyming lyrics. The word rapping in itself means "spoken or chanted rhyming lyrics". The foundation of rap goes back to the early 1960's when artists in both New-York and Jamaica mixed existing music with rhyming lyrics. The best rap tunes are the ones the that combine the basic elements of namely Context, Flow, Content and Delivery to the requisite amount. The late 1980's was the time when a number of individuals and groups took centre stage in the western world. Stylistically, rap music occupies a grey area between prose, poetry, singing and speech.

The very nature of rap being free-flowing and fast paced, makes it harder to analyse. The primary objective of Semantic Textual Similarity (STS) is to determine word relationships between different words in a corpus. In the words of (Agiree et al., 2012), it is the measure of equivalence between the sentences. STS is the primary component in many tasks like text reuse detection (Paul Clough et al., 2002), Twitter Search (Sriram et al., 2010), paraphrase recognition (Dolan et al., 2004). The primary goal of STS is to create a framework that can determine various aspects of similarity between texts. The importance of textual similarity is growing by the day with rapid growth in Artificial Intelligence and Natural Language Processing which is a result of a greater need to determine similarity between sentences or short text sequences which help in data classification among other tasks. There are a number of different methods available to determine textual similarity like information retrieval vector space model (Meadow et al.,), in which the text is modelled into a "bag of words" and represented as vectors. In the bag-of-words methodology, the similarity between two texts is determined as the cosine of the similarity of the vectors. The other method is to provide context for the short texts and enhance their vectors using the words in the snippets (Sahami and Heilman, 2006). The third method is the use of Google's Word2Vec framework. The continuous bag-of-words and skip-gram architectures for the computing vector representations of words. It is a shallow two layered neural network architecture that learns the distributed representations of words. Word2Vec does not require trained data for processing. With Word2Vec(W2V), a very large dataset can be easily classified by either using the Continuous-Bag-Of-Words or the skip-gram methods. Other recurrent neural network techniques are available for use but they take a long time to train their models. If a W2V network is given a large amount of data it produces distinct characteristic vectors that have words with similar meaning and context bundled into a single cluster, with a number of such clusters formed. W2V also does not require labelled data.

II. RELATED WORK

Sematic Textual Similarity is the technique of determining the similarity between a set or corpus of words using some similarity measure. One of the methods of determining similarity is the Semantic Textual Similarity System, here the Bag of Words are also called as the continuous bag of words method in determining the similarity

In the implementation of the Semantic textual similarity the authors describe three text similarity systems developed for the SEM 2013 STS shared task. Each of them had a common word similarity attribute that combined LSA and WordNet knowledge. The three approaches used are:

- Bag of Words.
- Semantic Word Similarity Model.
- Align-and-Penalize Approach.

Abhay Kashap et al., describes UMBC's system developed for the SEMEval 2014 tasks on Multi-lingual Semantic Textual Similarity (Task 10) and Cross-Level Semantic Similarity, which is an extension of the UMBC Semantic Textual Similarity system and describes at length the different approaches to determine the word similarity [1].

Aliaksei Severyn et al., describe a approach different from the majority of approaches, where a large number of pairwise similarity features are used to learn a regression model, this

978-1-7281-4142-8/$31.00 ©2020 IEEE

model directly encodes the input texts into syntactic/semantic structures. The system proposed rely on tree kernels to automatically extract a rich set of syntactic patterns to learn a similarity score correlated with human judgements [2].

Danilo Croce et al., describe a method of using SV regression which has become one of the key methods to determining similarity using SV regression on the UNITOR system, an entrant to the *SEM 2013 shared task on Semantic Textual Similarity (STS). The task is modelled as a Support Vector (SV) regression problem, were a similarity scoring function is determined between text pairs from examples. The proposed approach has been implemented in a system whose aim is at delivering high applicability and robustness, and reduce any contingency of over-fitting of datasets [3].

Features play a role in the determination of similarity. Irrespective of the methodology followed in the determination of similarity, feature selection plays a vital role. Erwin Marsi et al., the paper Combining Strong Features for Text Similarity forges the task accomplished at NTNU as part] of the *SEM'13 shared task on Semantic Textual Similarity that adopts a methodology where the concepts of shallow textual, distributional and knowledge-based features are determined using support vector regression model [4].

Context determination is also an emerging area of research in the semantic similarity field. One of the methods of determining context or aspect is the use of sematic labels. Pedro P et al., describes NILC USP that participated in the Semeval 2014: Aspect Based Sentiment Analysis task. This system uses a Conditional Random Field (CRF) algorithm for extracting the aspects mentioned in the text [5]. Our work added semantic labels into a basic feature set for measuring the efficiency of those for aspect extraction.

Context determination is fast becoming a very important area of research. Once ascertained, context or aspect will allow large organizations to effectively identify consumer patterns. In the paper Aspect and

Label identification in Consumer Review the authors describe a method to determine sentiments and user aspects from reviews. The system was proposed was for the SEMEval 2014 shared task. Here concepts like lexicons and word clusters are used.

Databases are highly important assets to any organization. One key problem with large databases is the querying for relevant data. It is in this scenario of finding relevant data that text similarity thrives. William W. Cohen et.al in the paper Integration of heterogeneous databases without common domains using queries based on textual similarity the authors make use of all databases that are currently available and map real world aspects like name, place, contact number etc directly onto a database which may become rudimentary is many cases [6]. The proposed methodology describes heterogeneous database integration using by making use of natural language constructs. The paper also describes a concept called Whirl and also implementation that is used to determine word similarity which intern reduce query time.

Sentence similarity plays a key role in many different textual research and applications and many areas like text mining, Webpage retrieval and dialogue systems. This paper describes a method along with an algorithm for processing sentences of small length by making use of word order information and sematic information.

Semantic similarity between words is becoming a generic problem for many applications of computational linguistics and artificial intelligence. This paper explores the determination of semantic similarity by a number of information sources, which consist of structural semantic information from a lexical taxonomy and information content from a corpus. To investigate how information sources could be used effectively, a variety of strategies for using various possible information sources are implemented.

III. Methodology

A. Existing System

Determining semantic textual similarity has been a subject of research in natural language processing since the beginning of early 1990's resulting in a number of different researches making varied research implementations of for determining such similarity or a plethora of different domains and datasets. Such research experiment has included fields as varied as medical healthcare. Rost et al., to trivial fields like movie reviews. Determining text-similarity also plays an important role in a number of different applications. like recommendation systems and so on [7]. Over the years, a number of different methodologies have been implemented for varying fields, but at the core of every text similarity implementation are three predominant methods of implementation

a) The first method is the use a Word Space model implementation to determine word similarity or text similarity, the words in the dataset are modelled into a so called "bag of words" and these bags are intern represented as word vectors. Here the similarity is determined using the cosine values or distances between the words or text within the vectors.

b) The second method used to determine textual similarity uses complex text similarity measures to determine the syntactic, semantic and lexical features of the data. The processed data is then supplied to a classifier which assigns weights to the features extracted by fitting the features to the training model [8].

c) The third method available to determine text similarity, it is initially assumed that the sentences or data currently under consideration are syntactically and semantically equivalent. With the equivalence assumption, having been made the sentences are aligned based on the corresponding assumption. In this method, the alignment quality or in other words the extent to which a given sentences are aligned with one another is used as the similarity measurement tool.

B. Implementation

As RAP is a very varied and highly content and context-oriented form of music, the best method suited for determining textual similarity in rap data is to use the "Continuous Bag of Words" or CBOW classifier as the word occurrence plays a very key role in similarity determination. The Continuous bag of words architecture is a very simple representation of words used in natural language processing and mainly in information retrieval. A given text or document is represented as a clustered bag word. The clustered bag does not give regard to any grammar or word occurrence order, keeping track of only the number of times of occurrence of a particular word. The most renowned use of the Bag-Of-Words classifier is in document classification where the occurrence of each word is used as a feature.

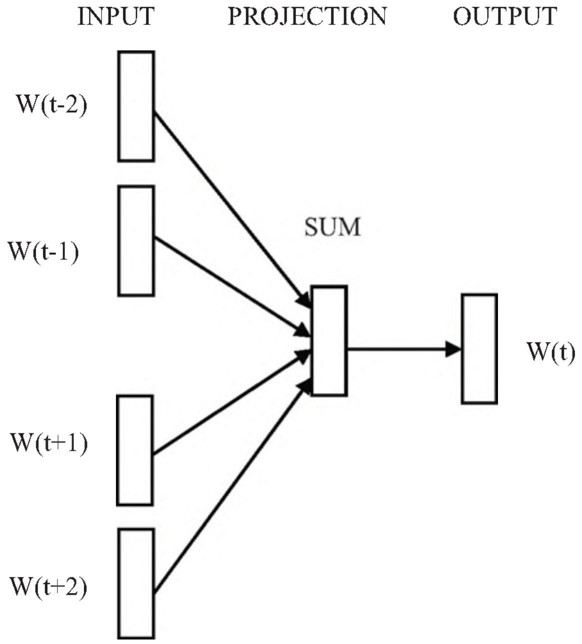

Fig. 1. Word2Vec word aggression diagram.

In the continuous bag-of-words model the projection layer is shared by all the words in the given data, and all the given words are projected in the same position after their corresponding vectors are averaged. The CBOW uses a continuous distributed representation of the data. Modelling the data using the traditional CBOW method is highly time consuming when a large volume of data is taken into consideration and the genre of data used in this scenario highly voluminous and making the use of the traditional method is inefficient. Another drawback is to keep track of the both the history words and future words when the size of data is large. In order to satisfy the voluminous nature of the dataset, Google's Word2Vec framework is applied in the implementation. Word2Vec is a deep-learning method where the meaning associated with data play a primary role in similarity determination.

Word2Vec takes as input a text dataset and produces word vectors as output. Word2Vec first builds a vocabulary from the text input and then learns vector representations of the words in the dataset. After the learning process in complete a word vector file is created which can be used as features in further processing in determining similarity.

C. Data Collection

To produce effective results using Word2Vec, a very large data corpus is required. The Word2Vec documentation describes that with increase the size of data there is an increase in the efficiency of the results. The data used was extracted from various internet sources offering data and metadata for usage. The subset of the dataset used consists data of 50,000 rap genre music instances with each of the instance containing the following features:

- The first field is the track which describes the name of the by which the particular record is released. The track name is synonymous to the name by which the audiences identify the record.
- The dataset contains the artist field which gives the name of the primary artists of the track. The artist field contains the detail description about all the primary performers on the track. It may contain multiple artist names based on the number of artists featured in the track. From the perspective of the dataset the artist details may be repeated multiple times.
- The primary and most important field in regards to the dataset in the lyrics attribute. This field contains the entire lyrics and not just a snippet of the lyrics. The featured attribute gives information about other artists involved in the track excluding the primary artists of the track.

D. Pre – Processing

Pre-processing involves the process of preparing the data for the implementation of the technique. To prepare the data for processing, the data must first be read into the required format. The data read in the proposed methodology is performed using the panda's framework for python. Once the data is obtained in the required format, the data must be cleansed to remove any noisy and invalid data. Here data is cleansed by removing all html tags and other delimiters. Stop-word removal is avoided as the algorithm relies on the context of each sentence to produce better word vectors. One of the primary constraints of Word2Vec is that is requires sentences with each of these sentences being a list of words, i.e. Word2Vec requires a list of lists.

E. Model Training

Training the data model involves creating a model from the parsed data that is, the list of lists. In Word2Vec, there are number of different parameters to be considered to train model:

- Architecture: There are two architecture available namely The Continuous Bag of Words and the skip-gram architecture. The skip-gram architecture is slower whereas the CBOW is faster. The method proposed uses the CBOW architecture.
- Training Algorithm: The hierarchical softmax algorithm is selected for use.
- Word Vectors Dimensions: Generally, the with the increase in the number of dimensions there is an increase in the quality of the model, but not always. Here a word dimensionality of about 300 words has been selected.
- Context size: this involves determining the number of words of context that should be selected. a value of 10 works well with hierarchical softmax.
- Word Count: Determining the size of the vocabulary plays a very important role in determining the overall quality of the data model. Limiting the number of the words in the vocabulary is evidently necessary for any model. Modest values for vocabulary ranges from anywhere between 5 to a hundred. Here a vocabulary size of about 30 was used. Here a value of 30 means that if a certain word does not occur at least 30 times throughout the data it is ignored.
- Sub-sampling of frequent words: Google documentation for proposes the use of values

anywhere between 1e-3 to 1e-5. It was found that the use of 1e-5 increasing accuracy.

F. Result and Discussion

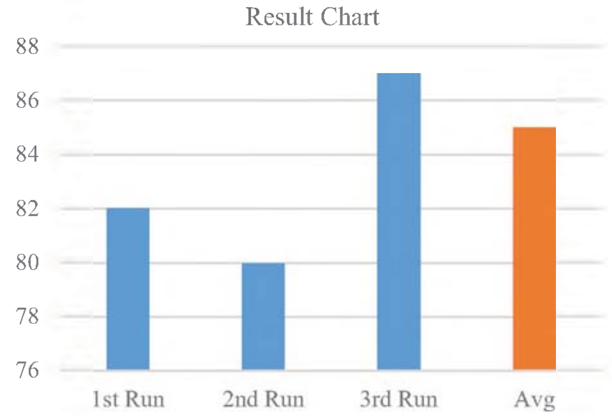

Fig. 2. Result across 3 runs and Average.

The method proposed in this paper was conducted across 3 runs. The best score being obtained in run 3 with a value totalling to 87% in accuracy. The above figure depicts the score across 3 runs, with an average score of 85% in accuracy.

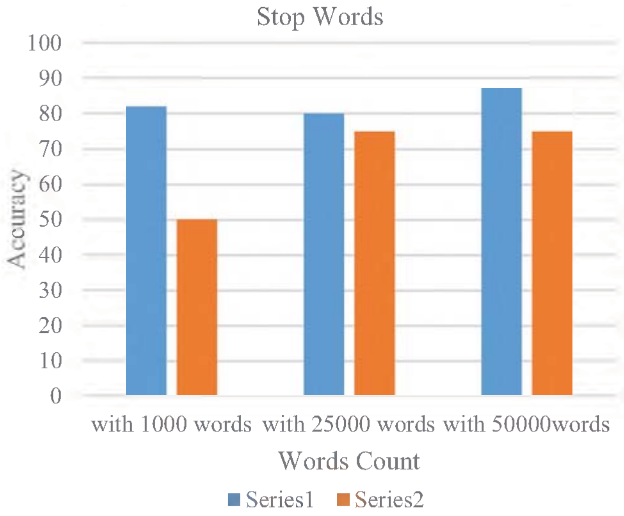

Fig. 3. Results with different data size.

One of the primary advantages of using google W2V is its ability to handle large amounts of data. In fact, as mentioned before, as the amount of data increases the accuracy of the output increases. In W2V the accuracy of result is directly proportional to the size of data. This property of Word2Vec is displayed in the above graph. In order to verify this property, Word2Vec was fed with different sizes of data across 3 runs. In the first run, a dataset containing 1000 words was input, which intern returned an accuracy of about 50%. In run 2 a dataset of size 25000 was input resulting in an output with accuracy 67%. In run 3 the entire dataset of size 50,000 was input and resulted in an output with accuracy of 87%. Therefore after 3 runs of execution, and average score of about 84% was obtained.

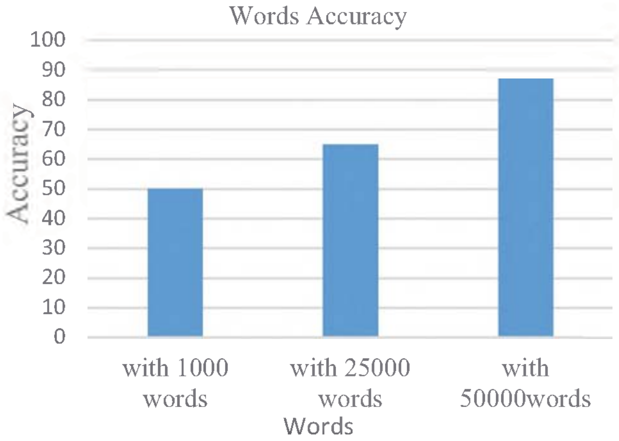

Fig. 4. Results with and without Stop words.

Stops words play a very important role in determining text similarity using CBOW in Word2Vec and in determining similarity in general. It is with the help of these stop word collections that some researchers filter data while others use them to determine context. With regards to rap lyrics, these stop-words are used to determine the context of the text. In order to substantially prove the importance of stop-words in determining similarity in rap lyrics, a test of three runs was additionally conducted but these runs were conducted after the removal of stop-words from the data-set and the results were plotted in combination with previously obtained results. As depicted in figure 4, the results in all the three runs conducted without stop-words being considered for contextual equivalence resulted in poor scores. Thus, it is effectively proved that stop-words are indeed required in determining similarity of text using W2V.

IV. Conclusion

Semantic Textual Similarity is an emerging area of research in the field of data mining and artificial intelligence. One of the key areas of implementation of this textual similarity is to ascertain textual similarity in music. The current research work makes use of lyrics from a large volume of data to determine similarity between the words or sentences in the corpus. Within the music cluster, the subset of Rap has the highest amount of lyrical content among all the musical forms and hence a large amount of research focus is being invested is this rap. Here, two primary features namely lyrics and track have been used to determine the similarity among data in the dataset. The dataset is of size 50,000 records of rap music tracks. The Continuous Bag of words classifier is used in conjunction with the Word2Vec framework provided by google. The outcome of the research resulted in a highly optimized word similarity model with an average accuracy of 85%. This model can further be used to build effective recommendation systems RAP music.

Acknowledgment

The authors would like to thank Prof. Joy Paulose, Head, Department of Computer Science, Christ University for his undying support through this entire process. He has been a source of inspiration throughout the entire process.

REFERENCES

[1] Abhay Kashyap, Lushan Han, Roberto Yus, Jennifer Sleeman, Taneeya Satyapanich, Sunil Gandhi and Tim Finin, The 8th International Workshop on Semantic Evaluation 2014, Meerkat Mafia: Multilingual and Cross-Level Semantic Textual Similarity Systems.

[2] Aliaksei Severyn, Massimo Nicosia and Alessandro Moschitti. iKernels-Core: Tree Kernel Learning for Textual Similarity. Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity 2013, Volume 1,

[3] Alejandro Riveros, Maria De Arteaga, Fabio González, Sergio Jimenez and Henning Müller. Comparing Metamap and T-mapper for Medical Concept Extraction. Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity 2014, Volume 1.

[4] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas 2010. Short text classification in twitter to improve information filtering. Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pages 841–842.

[5] Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In Proceedings of the 20th international conference on Computational Linguistics, COLING 2004 . Association for Computational Linguistics

[6] Danilo Croce, Valerio Storch and Roberto Basili, UNITOR-CORE TYPED: Combining Text Similarity and Semantic Filters through SV Regression.. Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity 2013, Volume 1

[7] Erwin Marsi, Hans Moen, Lars Bungum, Gleb Sizov, Bj¨orn Gamb¨ack, Andr´e Lynum, NTNU-CORE: Combining strong features for semantic similarity. Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity 2013, Volume 1.

[8] Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. Association for Computational Linguistics. Semeval-2012 task6: a pilot onvsemantic textual similarity. In Proceedings of the First Joint Conference on Lexical and Computational Semantics, pages 385–393.

[9] Lushan Han, Abhay Kashyap, Tim Finin James Mayfield and Jonathan Weese, Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, Volume 1, Semantic Textual Similarity Systems.

[10] Paul Clough and Robert Gaizauskas and Scott S.L. Piao and Yorick Wilks. METER: MEasuring TEx Reuse. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics 2Page 152-159

[11] Sahami, M., and Heilman, T. 2006. A web-based kernel function for measuring the similarity of short text snippets. In Proc. of the 15th Int'l World Wide Web Conference.

[12] William W. Cohen, SIGMOD '98. Integration of heterogeneous databases without common domains using queries based on textual similarity. Proceedings of the 1998 ACM SIGMOD International conference on the Management of data, Pages 201-2012,

[13] Wes McKinney, Proceedings of the 9th Python in Science Conference 2010, 51-56, Data Structures for Statistical Computing in Python.