National College of Ireland

A Natural Language Processing Approach for Musical Instruments Recommendation System

MSc Research Project
Data Analytics

Abhishek Dahale
Student ID: x17170311

School of Computing
National College of Ireland

Supervisor: Dr. Anu Sahni

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Abhishek Dahale
Student ID:	x17170311
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Anu Sahni
Submission Due Date:	12/08/2019
Project Title:	A Natural Language Processing Approach for Musical Instruments Recommendation System
Word Count:	6541
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	10th August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Natural Language Processing Approach for Musical Instruments Recommendation System

Abhishek Dahale
x17170311

Abstract

A number of significant researches have been made in the field of data analytics for recommendation systems and are able to successfully recommend the relevant and related content of the user's interest. Variety of domains like books, restaurants, web pages, movies, e-commerce, etc. make use of recommendation systems, thus playing an important role in recommending the items of user's interest. The outcomes produced by these systems always have the scope of improvement over the results they currently provide which would lead to the user satisfaction . This research strengthens the use of the recommendation system by improving the accuracy of the results provided. In the recent years, the demand for musical instruments is increasing with the growing interest of people in opting music as a career or an extra-curricular activity. Hence, this research focuses on building a recommendation system for musical instrument, that will generate top 3 recommendations based on the search query entered by user. This search query contains details in form of technical specification, description, sentiments, and brand related to musical instrument. A Natural Language Processing approach is followed in order to generate recommendations by training the models: Dov2vec and Latent Semantic Analysis(LSA) using the dataset from Amazon . Further, the cosine similarity between a chatbot search query is matched with the average of the cosine similarity of LSA and Doc2Vec document embeddings to generate recommendations.The cosine similarities of the results obtained for top 3 recommendations from the experiments performed was 0.825319, 0.807372 and 0.789747.

Keywords- Natural Language Processing(NLP), Doc2Vec (Paragraph Vectoring), Recommendation System, LSA(Latent Semantic Analysis)

1 Introduction

Internet shopping is booming in todays world of e-commerce. For shopping, a person often finds the easiest and most convenient way of doing it. Thus, e-commerce has brought all the online buyers to a single platform creating an accessible way of shopping. According to the survey of the global retails sales in e-commerce activities all over the world, the share accounted in 2017 was raised to 17.5% which was just 7.4% in 2015 [1].Shopping over the internet has a lot of advantages which brings convenience in shopping, but there are certain disadvantages and limitations that need to be considered. The product we

[1]https://www.statista.com/statistics/534123/e-commerce-share-of-retail-sales-worldwide/

are buying and the product what is delivered doesnt always meet our expectations, often leading to disappointment. As the requirement of the customers keeps changing day by day, the enhancement of existing functionality of these systems has become an important factor. Due to urbanization, the shopping behavior of the person has lead to a lot of changes. The number of digital buyers worldwide using e-commerce platform in 2018 was 1.79 billion which are expected to be 2.14 billion in 2021[2]. This shows the increase in huge demand for online shopping in the near future. Thus we can consider one such example of an increase in the demand for music instruments worldwide. According to the Toys & Hobby report 2019 for music instrument, the revenue generated in the musical instruments amounted to 32,590 million USD which is expected to grow by 5.7% per year[3]. Increase in the demand is mainly due to the involvement of music a hobby or as a career or an extra-curricular activity . Choice of the instrument for musician depends on the tang of the music preferred, which differs from person to person. Also the factors involved while deciding the prices of these musical instruments that include the technical specifications, brand, quality, and type. While buying these instruments online from e-commerce websites, a lot of problems need to be faced. Also, the product ordered online by just browsing the name or the category would lead to dissatisfaction and waste of time and money. Thus, this research follows the natural language processing approach to build a musical instrument recommendation system. A chatbot interface provided will help the customer to describe what type of musical instrument is required. This description would be in the form of feelings, mood, description, specification, and brand, that will help the user to get relevant recommendations.

1.1 Relation between recommendation system & data analytics

About 2.5 quintillion bytes of data is generated in a single day which is expected to be 1.7Mbps in 2020 as reported by DOMOs[4].To manage this gigantic data, information filtering has become an important task due to this large amount of information generated every day(Hanani et al.; 2001). Therefore, a key method that can be considered in information retrieval and filtering of a large amount of data is a recommendation system (Davidson et al.; 2010). As online shopping has opened up a lot of options for users, the need to process this information has also increased. Thus a recommendation plays a vital role in (Schafer et al.; 1999a). This research would help to filter a large amount of data and generate recommendations of user's interest.

1.2 Motivation

As recommendation systems are successful in recommending the items of user's interest, a number of researches have been made. The relevant content presented to the user is obtained by filtering a large amount of information. The satisfaction of the user by provided recommendations shows the accuracy of the system. While shopping, as number of options are available for us, we often get confused. Thus, the motivation behind building this recommendation system is, when a person intends to buy a musical instrument from any e-commerce web site, due to number options, it makes user confused. Often a product received leads to disappointment. Thus to avoid this, a recommendation system

[2]https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/

[3]https://www.statista.com/outlook/19020000/100/musical-instruments/worldwide?currency=usd

[4]https://www.domo.com/solution/data-never-sleeps-6

will help in customer satisfaction and generate revenue to the business by providing recommendations. In this research, A Natural Language Processing approach is followed to generate recommendations by training the models: Latent Semantic Analysis(LSA) and Dov2vec. A chatbot interface will help to communicate with these models. This type of system would help in solving the challenges faced by customers including new beginners and music professionals, increasing customer satisfaction and also maximizing the business revenue.

1.3 Research Question

The research question addresses the problems faced on the e-commerce platform by musicians or beginners or any person interested in buying a musical instrument.

Research Question: *"How the Natural language processing approach (Using models Doc2Vec and Latent Semantic Analysis) can be used for generating top 3 relevant recommendations of the musical instrument based on the description entered by the user in the form of sentiments, description, technical specifications and brand using chatbot?"*

The report is organized as follows: Section 2 contains critical reviews on the literature work related to the recommendation system using different machine learning approaches. Section 3 presents the approach used and how the models will be used to answer the research question. Section 4 presents the design specification that gives an overview of the architecture of the recommendation system. Section 5 presents the overall implementation of the research project. Section 6 discusses the evaluation and results. Section 7 includes discussions related to the research project. At last, the conclusion and future work in discussed in Section 8 followed by acknowledgement and references.

2 Related Work

The amount of digital information is growing exponentially leading to overloading of data and has become challenge for web users to search the required information. The problem of information availability is solved by the search engines like Google but the challenge of personalizing the contents of user's interests still persist. Thus a need to build recommendation system has increased for overcoming these challenges faced by internet users globally(Isinkaye et al.; 2015). Thus, to filter a large amount of information and recommend according to user's preferences and interest(Pan and Li; 2010), a recommendation system can be used. Pu et al. (2011) employed research on the success of recommendations from the user's point of view. A framework called ResQue, Recommender Systems Quality of Users Experience was used to evaluate the performance. This evaluation was based on measures that include qualities like the usability of the system, interaction quality, users satisfaction and users behavior etc., that helped in explaining motivation and the users experience behind adopting recommendation system.

The recommendation systems are shaping E-commerce platforms. It learns customer behavior and recommends them with interesting and valuable products (Schafer et al.; 1999a). Websites such as Amazon, Levis, Moviefinder.com have their own way of generating recommendations. For example, Amazon[5] recommends user items that have different features like Customer who viewed this item also viewed, Customer who bought this

[5]www.amazon.com

item also bought, Book Matcher, Amazon Delivers, etc. Match Maker feature is used by MovieFinder.com[6] where recommendations are generated based on theme, genre or mood. The WePredict feature of it recommends the movies based on users past interests. (Schafer et al.; 1999b).

Recommendation system can be classified as content-based filtering, in which the items are recommended by understanding the user behavior (Pazzani and Billsus; 2007) or collaborative, where recommendations are based on reviews (Schafer et al.; 2007). Different data sources are used by different recommendation system for balancing factors like accuracy, novelty, and stability (Bobadilla et al.; 2013). This estimate is then further used for recommending items of user interests. Pazzani and Billsus (2007) used different user modeling algorithms which involved Bayesian classifiers and Decision Tree in Content-based filtering. For providing accurate results, they introduced a new technique of relevance feedback, that allows to adjust model on the basis of data source. In Collaborative filtering, other user's opinions are considered(Herlocker et al.; 1999).

Mooney and Roy (2000) developed a system for recommending books using text categorization and information extraction based on content-based filtering, which helped them in achieving better results for recommendations. A prototype, Learning Intelligent Book Recommending agent(LIBRA) was developed. The data source used for this research mainly involved extract from Amazon.com web pages. Bayesian Learning Algorithm was used to learn profiles of users and generate recommendations for books according to their rankings. LIBRA performs simple pattern-based information extraction, that generates meaningful insights from data. Based on the probabilistic binary categorization, the book is rated as negative(1-5) or positive(6-10). Thus using Naive Bayes, profiles are learned and the books with top score are recommended to the user, which proved LIBRA to be significant and consistent.

Lang (1995) proposed NewsWeeder, a Netnews-filtering system and described how it learns user profiles based on user ratings and reading the articles. It uses both, content-based and collaborative filtering. MDL(Minimum Description Length), a machine learning technique helped them to increase the percentage of a user interested from 14% to 52%. MDL outperformed over TF-IDF by 21% and helped them to achieve precision of 44%.

Nath Nandi et al. (2018) used Doc2Vec algorithm for news recommendation system in Bangla language. Document embeddings were generated using Doc2Vec model to capture the semantic similarity between the documents. They performed quantitative and qualitative experiments to evaluate Doc2Vec model by comparing it against LDA(Latent Dirichlet Allocation) and Latent Semantic Analysis (LSA). Nath Nandi et al. (2018) proved that Doc2Vec outperformed over LDA and LSA. The accuracy achieved with helps Doc2Vec was 91% while that of LSA and LDA was 84% and 85% respectively.The accuracy of these models was very high which helped in generating recommendations but sentiments were not considered while recommending news by Nath Nandi et al. (2018). As our research involves sentiments to be taken into account, using this model along with sentiments about the instruments user is interested in, would help to boost the overall performance.

Pazzani et al. (1998) developed Skyskill & Webert, a software agent, that learns user interested pages. It learned user profiles by analyzing the information on every page and recommended the links of the user's interests. They constructed a web search engine called LYCOS that helped user in searching interested pages. Pazzani et al. (1998) found

[6]www.moviefinder.com

that recommendation generated with help of Nave Bayesian Classifier achieved 77% accuracy. The Nearest neighbor and backprop were found to be 75% accurate with ID3 to be least one i.e. 70% accurate. They also performed experiments using different Information Retrieval algorithms using TF-IDF to perform classification. Thus, recommendations for web pages were generated by learning profiles of user. Similarly, Armstrong et al. (2003) developed a web pages recommender system called as Webwatcher. They used different training methods viz. TF-IDF with cosine similarity and Wordstat-for making predictions of links.

Gomez-Uribe and Hunt (2016) from Netflix Inc explained different algorithms that are used by Netflix. A Top N Video ranker produces best and top recommendations by searching the entire catalog. PVR, a personalized video ranker algorithm personalizes order of videos that differ from person to person according to the genre. A Trending Now algorithm uses a trending ranker to recommend the trending movies. A video ranking algorithm, Continue Watching Ranker is used for ranking the videos. Thus, all the above-mentioned algorithm generates a personalized page using a page generation algorithm based on the interest of user. Gomez-Uribe and Hunt (2016) also mentioned that the system is built on Machine Learning and statistical techniques which includes supervised (classification and regression) and unsupervised(clustering and compression) approach. However, the accuracy of this algorithm was not discussed by them.

Lund and Ng (2018) proposed a movie recommendations system. They used deep learning for verifying the novelty by comparing it with a collaborative-filtering technique: matrix factorization and KNN(k nearest neighbor). They used Data sources from MovieLens to generate recommendations where the approach used by them has better results as compared to models mentioned above. Results obtained by RMSE for KNN and Model-based approach is shown in the below table

	User-user KNN	**Model-based**
Train	NA	0.4209
Test	11.6715	0.3544

Alspector et al. (1998) discussed feature and clique based user models for recommending movies. In a feature-based approach, a feature of movies is extracted using CART (Classification and regression tree) and neural network and In clique based approach, users are categorized on the basis of similar cliques. Alspector et al. (1998) compared efficiency of the above feature and clique based models, of which clique based was used when information of the user was available, whereas feature-based approach was used in case where there were no user ratings.

Linden et al. (2003) from Amazon.com proposed of having personalized online store to each customer using recommendation algorithms. It uses item-to-item collaborative filtering and filters similar items . It finds the items that match with items user has bought and rated and recommends them to the user. By using the iterative algorithm, it calculates similarity using the cosine measure for the related products. Linden et al. (2003) proved how item-to-item collaborative filtering provided high-quality recommendations as compared to traditional collaborative filtering but this discussion did not involve the accuracy of algorithms. As personalized store is created based on purchases of customer, it is required to learn the behavior of user to generate recommendations to new users. The recommendations generated did not consider the user's sentiments. This research involves using chatbot that will recommend items considering user's sentiment.

Tan et al. (2008) developed a system which helped in generating recommendations of courses for e-learners, known as E-learning recommendations. User-based collaborative filtering was used in recommending the related contents. The recommendation were generated by forceful rating of courses by learners. The rating scores were transformed using Learners Rating Matrix and similarity between like-minded learners and target learners was calculated to form a proximity-based neighborhood. Neighborhood formation algorithm and Proximity measure were used, where Pearson Correlation was used to calculate the proximity. Here in this type of recommendations, the user explicitly needs to rate the course and then the recommendations are generated by learning the user behavior, which becomes a drawback for the users using it for first time. The recommendation system build in this research would consider the drawback mentioned above while generating recommendations,that would include user's sentiments.

Lu (2004) developed a similar personalized e-learning system, PLRS (Personalized Learning Recommendation System) by using fuzzy rule matching developing a multi-criteria student requirement analysis model. Lu (2004) also discussed the inefficiency and incapability of DTs over Fuzzy set technique. Top-N recommendations were generated by evaluating a similarity measure. Lu (2004) did not how efficient fuzzy rule matching approach is, in generating recommendations.

Ma et al. (2017) developed a recommendation system for courses based on semantic similarity for students. The course content was modeled in the form of documents. The course vector was calculated using Doc2vec for each document and the courses were recommended to students based on cosine similarity. Students were just able to get the recommendation of the courses as per the search query entered by them. Ma et al. (2017) did not consider sentiments of students context of what course they like or dislike. Considering the sentiments as a drawback, this research involves using chatbot that will recommend items considering user's sentiment.

2.1 Conclusion

Thus, considering overall related research above in the fields, to build an effective and efficient recommendation system, different machine learning and Natural language processing approaches were used to recommend relevant items to users of their interest. Recommendations systems are used in number of different domains including web pages (Pazzani et al.; 1998), e-learning (Tan et al.; 2008) , e-commerce(Linden et al.; 2003), books(Mooney and Roy; 2000), videos(Gomez-Uribe and Hunt; 2016),courses(Ma et al.; 2017), movies (Lund and Ng; 2018), news (Nath Nandi et al.; 2018), etc. But with reference to the above researches, a specialized and personalized recommendation system for recommending the musical instruments have not been developed yet. While searching any instrument we are interested, we just enter the name or description of what we need. To query these systems we cannot use our sentiments about the instrument to get relevant recommendations. However, to my knowledge, these types of systems seldom exist. Hence, considering the significance and the use of musical instruments, the recommendation system built will help musical instruments buyer to get exactly what is needed.

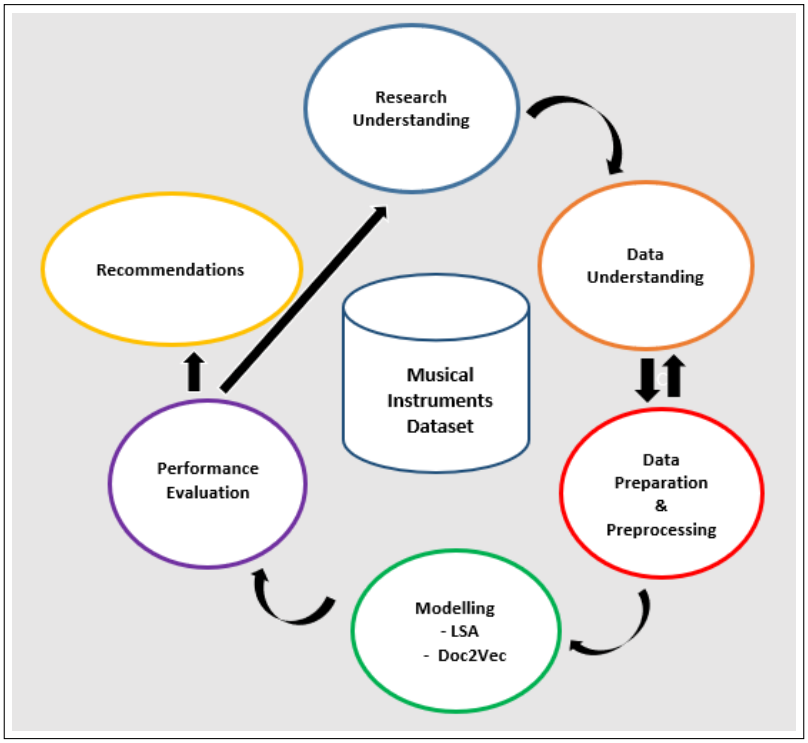

Figure 1: CRISP-DM

3 Methodology

As this research is based on building a musical instrument recommendation system, designing a flow of system efficiently is an important task. Different techniques involved in data mining process include CRISP-DM (Cross Industry Standard for Data Mining), KDD, CRISP-D, and SEMMA. After studying these techniques CRISP-DM was found to be efficient as it best fits the requirement of this research and is reliable, less costly and faster (Chapman et al.; 2000). The benefit of using CRISP-DM is we can modify it according to the application in research. According to the research requirement, the generalized CRISP-DM is modified as per requirements and is shown in Figure 1. The detailed outline of each step followed is discussed below:

3.1 Research Understanding

The initial phase of CRISP-DM involves understanding the research objective. The knowledge gathered is converted to a data mining problem. This phase consists of the overall planning of the project and discusses the aim of the research. Thus, the aim of this research is to build a music instrument recommendation system. The system will be able to recommend the musical instruments of the users interest based on his search query. This search query would contain a description in the form of feelings, emotions, description, and brand of the music instrument. A chatbot will be used to query a model with the search query, which will provide the top 3 recommendations to the user.

3.2 Data understanding

Collection of data is an initial task in performing research. Data availability becomes an important factor. As data required for this research contained the details in the form of description of the musical instruments, Amazon.com was used as the data source.A version of the dataset from amazon is provided at `http://jmcauley.ucsd.edu/data/amazon/links.html`. The dataset has information on 84,903 products related to musical instrument category which includes details like Title, Price, Image Url, Description, Sales Ranking, Categories, Brand, and Reviews. From the above-mentioned details, table 1 shows the attributes of data used in performing this research.

Table 1: Attributes of Dataset

Attribute	Description
Instrument_ID	Unique Id for each musical instrument
Title	Title of the musical instrument
Url	Url of the image to be displayed
Brand	Brand of the Product
Description	Detailed Description of the product
Reviews	Reviews posted for the product

3.3 Data Preparation & Preprocessing

After the data is chosen to be used for the further data mining process, there are different tasks that need to be performed on data. This dataset produced will form an input to the models used for recommending a musical instrument. As described in the above section, the attributes required for the research are selected. The next step involves cleaning the data. Cleaning of the data was done using R programming. As dataset contained 84,903 rows, it also included irrelevant data that was not required. This data was filtered using the categories, of which the data related to music instrument was filtered. All other rows having category apart from Musical Instrument were removed. Null values from the data were removed. The unwanted columns SalesRanking, Categories, Prices were removed. The final version of dataset contained the attributes mentioned in table 1 with 6124 rows.

As this research follows the use of Natural Language Processing, a large amount of text data needs to be processed before applying the machine learning models. The text data needs to preprocessed which involves tasks like changing data to lower case, removing stop words, tokenization, stemming, etc. The Figure 2 shows the steps followed for preprocessing of data. The detailed description of the steps followed is discussed below:

3.3.1 Lowercasing

Transforming the data to lowercase is the most effective form of data preprocessing and significantly helps in achieving consistent output. Hence, lowercasing is the best way to deal with sparsity issue. Table 2 shows the example of one of the rows from dataframe to which lowercasing is applied. A simple string lower() method in python is used to change the data to lower case.

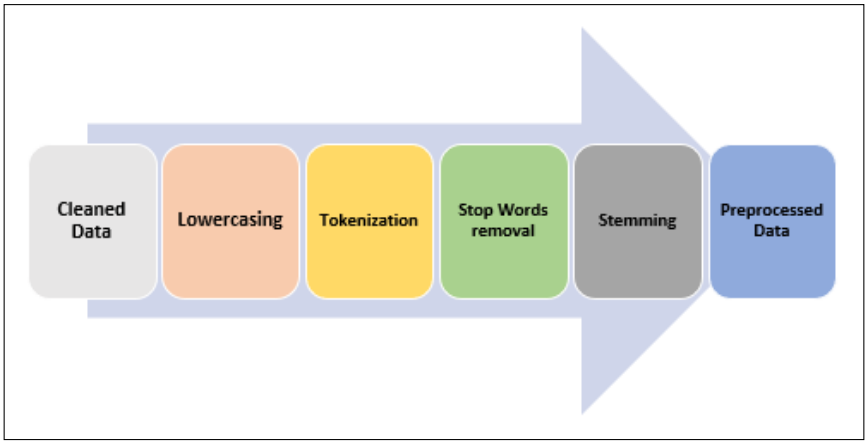

Figure 2: Data Preprocessing

3.3.2 Tokenization

After lowercasing, the next step followed is the tokenization where the sentences are divided into substrings known as Tokens. These tokens can be used to find the words in sentences. A simple regular expression based tokenizer RegexpTokenizer provided by NLTK was used which splits the text into punctuations and whitespaces[7]. Table 2 shows the example of tokenization applied over dataframe.

3.3.3 Stopwords Removal

Stopwords are generally a set of common words used in any language. For example is, the, a, an, etc. The importance of removing the stops words is that we can focus on the important words in context. It helps in reducing the number of features to be considered which helps to optimize the model. Table 2 shows the example of stopwords removal. A Natural language toolkit (NLTK) was used to load the stopwords and remove it.

3.3.4 Stemming

Stemming is the process of bringing word to their root form. Bringing to a root form just implies a canonical form of the original word. Stemming helps in standardizing vocabulary and dealing with sparsity issues, that matches all the variations of the words to bring the relevant recommendation. A SnowballStemmer provided by NLTK was used to perform stemming on the dataframe. Table 2 shows the example of stemming operation performed.

3.4 Modeling

As the data to be used as an input to the model is preprocessed and ready to use, two modeling techniques were applied for further data mining process. As this research follows the Natural Language Processing based approach, the recommendation system will be built with the help of the following models: Doc2Vec (Paragraph Vectoring) and Latent Semantic Analysis (LSA).As discussed by Nath Nandi et al. (2018), these models

[7]https://www.nltk.org/$_{m}$*odules/nltk/tokenize.html*

Table 2: Data Preprocessing

Preprocessing	Input	Preprocessed Text
Lowercasing	Yamaha's mouthpieces have an ideal weight	yamaha's mouthpieces have an ideal weight
Tokenization	yamaha's mouthpieces have an ideal weight	yamaha s mouthpieces have an ideal weight
Stopwords removal	yamaha s mouthpieces have an ideal weight	yamaha mouthpieces ideal weight
Stemming	yamaha mouthpieces ideal weight	yamaha mouthpiec ideal weight

are simple to use and provide high accuracy. The accuracy of the models successfully recommending the items to the user for LSA is 84% and that for Doc2Vec is 91%. These models were evaluated against cosine similarity and individual similarities of the models were averaged to get accurate and better recommendations. The functionality of these models is discussed below.

3.4.1 Latent Semantic Analysis(LSA)

Latent Semantic Analysis is a mathematical/statistical technique that derives the relation between expected use of words in the context of the sentence (Landauer et al.; 1998).It is a method of applying statistical computations to a large corpus for representing contextual- usage of words meaning. As each language has its own intricacies and nuances, it is difficult for a machine to capture its meaning. LSA comes into picture when words have the same spell but different meanings. LSA leverages the word context to capture the concepts that are hidden around words also called topics. So, simply mapping the words to the document prepared, here in our case the document contains the details like description, brand, and reviews of the instruments, won't really help. Thus to figure out the hidden topics behind the words, LSA is a simple technique to be used. A vector representation of these document would help in finding the similarity between the documents. Below steps are followed while implementing LSA:

- Generate a document-term matrix (m x n) that contains TF-IDF scores as shown below where,
  m: the number of documents that contains the details of the instruments.
  n: the number of unique words or terms.
- Using Singular Vector Decomposition (SVD), the dimensions are reduced.
- The matrix is decomposed into three other matrices. Suppose a matrix A (m x n) is to be decomposed using SVD. It will be decomposed as matrix U, matrix S and V^T represented as

$$A = U * S * V^T$$

Figure 3 represents LSA matrix where,
Uk is the document term matrix, where k is the length of the vector
Vk is the term-topic matrix which is the vector representation of the terms

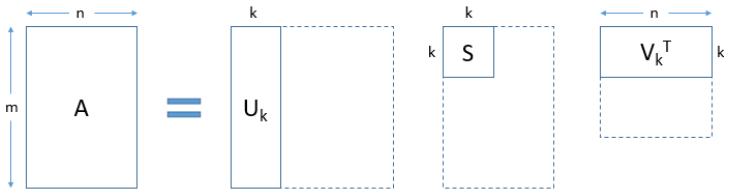

Figure 3: LSA Matrix

S is the diagonal matrix with non-negative numbers on diagonal used for scaling

- Thus, SVD generates vectors for every term and document. These vectors will be used to find similar words and document by calculation cosine similarity.
  Figure 4 gives the schematic overview of LSA.

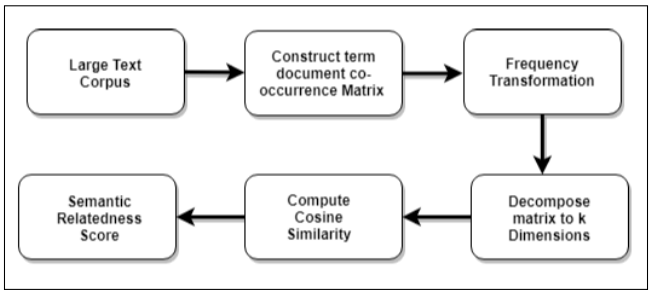

Figure 4: Latent Semantic Analysis

3.4.2 Doc2Vec (Paragraph Vectoring)

Paragraph vectoring or Doc2Vec is an unsupervised learning algorithm that generates a fixed-length representation of a variable-length text document (Le and Mikolov; 2014).It learns continuos distributed vector representation for the document, hence called an unsupervised framework. Doc2Vec predicts the words from each document using a dense vector. The advantage of using this model is that they can learn easily from the unlabelled data and also works very well labeled data. The Vector representation is useful in predicting the words in a document. These paragraph vectors are trained using backpropagation and SGD (Stochastic Gradient Descent). There are 2 algorithms used for learning the vectors. PV-DM(Paragraph Vector Distributed Memory Model) and PV-DBOW (Paragraph Vector Distributed Bag of Words).Le and Mikolov (2014) discussed of PV-DM to be effective and consistently better than PV-DBOW. Hence, for building this recommendation, PV-DM is used to achieve better accuracy by setting the parameter DM=1 during initializing the Doc2Vec model. Here, the word and paragraph vector are initialized randomly. The word vectors get shared among the documents while paragraph vector is assigned to every single document. Thus the paragraph and word vectors are averaged and passed to stochastic gradient descent and backpropagation is used to obtain gradient. Figure 5 shows the architecture of Doc2Vec for PV-DM. Here when we train

the Word vector W, the Document vector D gets trained as well, holding a numeric representation of the document whole document. The Input here is the Document ID vector and the Word vector. The Word vector has a dimension 1xV while the Document vector has dimension 1xC, where C is the total number of documents. W is the weight matrix of the hidden layer that has a dimension VxN and the weight matrix D has dimension CxN for the hidden layer.

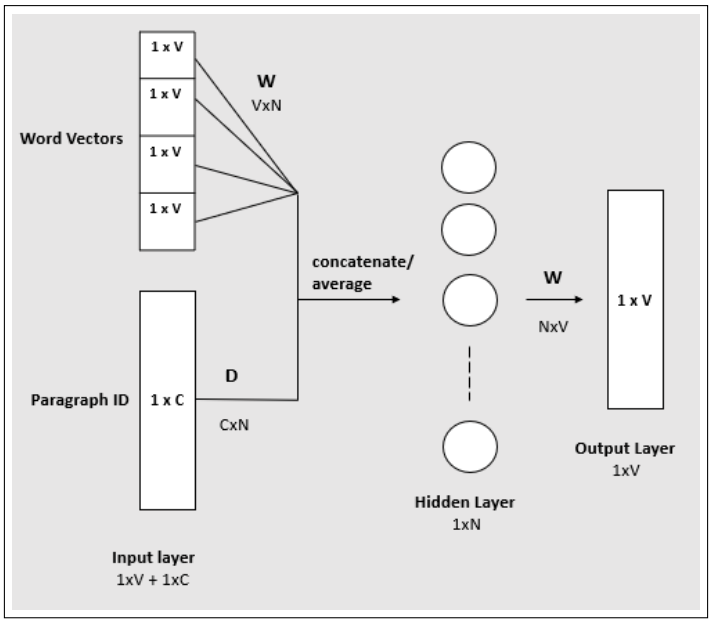

Figure 5: Doc2Vec (Paragraph Vectoring)

3.5 Evaluation

After the above models, LSA and Doc2Vec are successfully built, the next phase followed in CRISP-DM methodology is an evaluation of these models in terms of generality and accuracy of the model. Evaluation of the model assesses the degree to which the models build is successfully able to recommend the musical instruments of the users interest which meets the research objective (Chapman et al.; 2000) For this, the LSA model and Doc2Vec model will be ensembled. The models would present the details of the musical instrument and the search query entered by the user in latent space. Based on Cosine similarity, the details of the musical instrument and the query will be matched to recommend the musical instruments to the user. Evaluation of the models is further discussed in detail in section 6

3.6 Deployment

The deployment follows the last step of CRSIP-DM methodology, which includes deploying the data mining results in business. This also includes maintenance and monitoring of the applications if it becomes part of day-to-day business (Chapman et al.; 2000). The deployment of the music instrument recommendation system in a live environment to be used in day-to-day business will be the future scope of this research.

4 Design Specification

To develop an efficient recommendation system that would help in generating recommendations successfully, an architecture design shown in the figure Figure 6 is followed. This architecture presents the tools and techniques used in building a recommendation system for musical instruments. The architecture is divided into three blocks Data Persistence Layer, Business Logic Layer and Chatbot Interface. Data Persistence Layer describes the data source and data cleaning. Business Logic Layer contains pre-processing of data, document generation, Training and evaluation of models. Chatbot interface helps the user to communicate with the models in the form of text query and generate relevant recommendations.

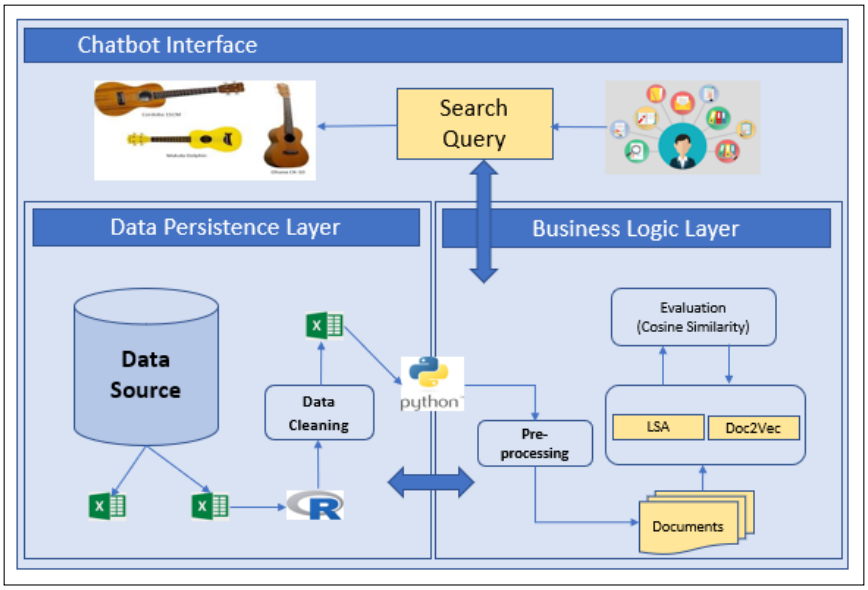

Figure 6: Architecture

The implementation of the server-side which helps in generating the recommendations is shown in Figure 7

5 Implementation

This section discusses the overall implementation of building the Musical Instrument recommendation system. It includes a detailed description of all the tasks carried out in this research to successfully recommend the instrument to the user. All the phases are discussed below:

5.1 Environmental Setup

The implementation of this project was performed on a 64 bit Windows operating machine with 8 GB of RAM. Python and R programming was used to build the recommendation system. Jupyter notebook provided by Anaconda framework was used for overall implementation. The latest version of Python 3.7.3 and for R 3.5.2 was used.

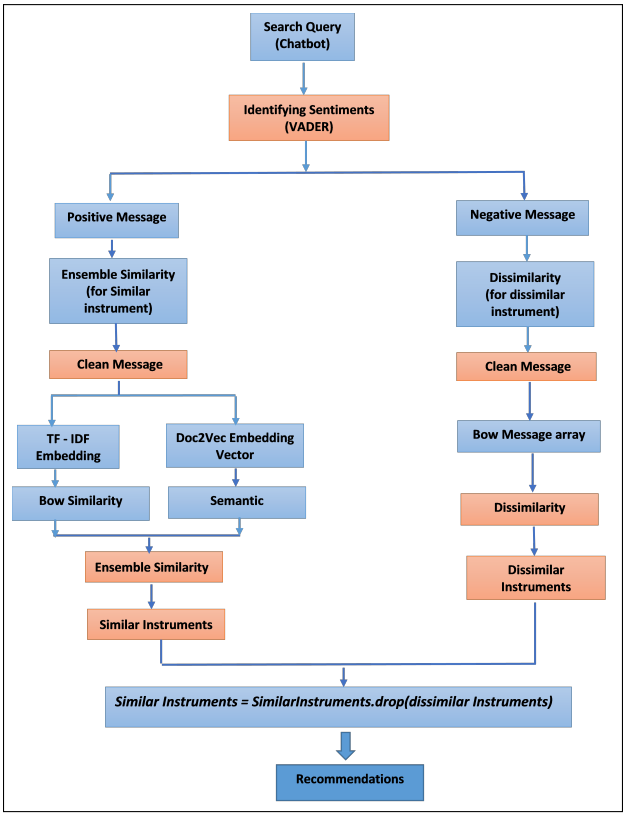

Figure 7: Implementation

5.2 Data Cleaning, Preprocessing & Transformation

Data Cleaning was performed using R programming. All the unwanted columns were removed and data was filtered based on the category as described in section 3.3.The cleaned data was imported in python using Pandas dataframe to perform further data preprocessing operations. Before the data was ready to be used as an input to the model, it was needed to be preprocessed. The data preprocessing tasks such as lowercasing, tokenization, Stopwords removal, and stemming are discussed in detail in section 3.3.Pickle module was to serialize and deserialize the Python objects. Thus, the data was ready to be used for the models.

5.3 Document Preparation

As this research follows the use of Doc2Vec and LSA algorithm, the input for this model is in the documented form. Therefore, to generate the document embeddings (Han Lau and Baldwin; 2016), the document is prepared that consists of the details related to a musical instrument including Description, Brand and Reviews. A new column is introduced in dataframe that consists of this document prepared.

5.4 Feature Extraction

5.4.1 TF-IDF

A document-term matrix was prepared to learn vocabulary. The sklearn feature extraction module was used to extract and build features from the documents [8]. Thus a TF-IDF vectorizer converts these document to a TF-IDF feature matrix. The n-gram range in the feature matrix was set as 1-4 to improve our match between the search query and the document of the music instrument created. Thus a TF-IDF matrix with dimension (6124 x 8313) was generated where 6124 is the number of documents containing the details of music instrument and 8313 is the number of unique terms in the document.

5.4.2 Latent Semantic Analysis

To perform the linear dimensionality reduction, truncated SVD (Singular value decomposition) was performed using the sklearn decomposition module [9]. Unlike PCA, this estimator does not perform the data centering for computing SVD. This works efficiently with scipy.sparse matrices. Thus, the truncated SVD is applied to the document-term/TF-IDF matrix and this transformation is known as Latent Semantic Analysis, as it transforms the matrices to the low dimensionality semantic space (Halko et al.; 2011). Thus the SVD model was dumped with the value of n_components=500, which is the desired dimensionality of the output data. SVD feature matrix was formed by picking 25 components. Thus an LSA embedding i.e. vector representation of a document is created to be used for further evaluation.

5.4.3 Doc2Vec

The input for Doc2vec model was prepared by forming a vocabulary that contained the Description, Review, and Brand of musical instrument. All the operations required for Doc2Vec was performed using Gensim, an open-source library for unsupervised natural language processing. The vocabulary prepared was tagged using TaggedDocument feature of genism module. The hyper-parameters used to create a Doc2Vec model is described in table 3. The model was built using different combinations of the hyper-parameters and tested with how relevant the recommendations are. The genism module build_vocab was used to build vocabulary from the tagged documents. The Doc2Vec model was trained using a different combination of epochs. The Doc2Vec model was dumped for further evaluation to give recommendations for musical instruments. Also, Doc2Vec embeddings were created by using Doc2Vec feature matrix.

5.4.4 Chatbot implementation

A chatbot is basically the computer program that emulates the communication with a human user to provide certain services, in our case, recommendations (Kucherbaev et al.; 2017). The user enters his search query for the musical instrument of his interest in the text box provided. The python widget framework ipywidgets was used to handle the events at client side. Different functions like widgets.Text(), widget.Button() were used creating the User Interface. Also, the event handlers on_submit and on_click were used to registers function on clicking and submitting the button. Thus, the important task

[8] https://scikit-learn.org/stable/modules/classes.htmlmodule-sklearn.feature$_e xtraction.text$

[9] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

Table 3: Description of Hyper-parameters

Hyper-Parameter	Description
vector size	Dimensionality of feature vector
epochs	Number of iterations on corpus
window	Distance between the current and predicted word
workers	threads used for training the model
dm	The training algorithm used
min_count	Ignores all words with total frequency lower than this
seed	seed for random number generator

of a chatbot to send the search query entered by the user at the back end and display the output in the form of recommendations. The detailed implementation of a chatbot is explained in the configuration manual.

5.4.5 Query Processing & Generating Recommendation

The search query entered by the user for getting recommendations is processed by a chatbot and passed to the server-side for further processing. The query is pre-processed such that it matches with the description to find similar instruments. As the search query consists of the sentiments of the user for the instrument, the first task is to get the sentiments of the query entered. For this, VADER sentiment Intensity analyzer (Hutto and Gilbert; 2015) from NLTK is used which divides the query into positive and negative parts based on the polarity scores of the query. Thus, giving us the part of love messages and hate messages from the original text query. The next step involves getting similar instruments using similarity scores based on the positive message in query and getting dissimilar instruments based on the negative message and removing them. The detailed implementation of how the relevant recommendations are generated is explained in Figure 7 of Section 4. Thus, the top 3 recommendations are plotted as the output of the chatbot with Title and Image of the musical instruments.

6 Evaluation & Results

6.1 Evaluation Metric

6.1.1 Term Frequency Inverse Document Frequency(TF-IDF)

TF-IDF is the statistical-based technique which is widely used in Information retrieval and search engines. It is used to weight the terms in the document. In the context of this research, to generate the document-term matrix to match the search query to get the relevant recommendations (Lahitani et al.; 2016). The TF-IDF is composed of two terms: The Term frequency (TF) and the Inverse Document Frequency [10] given by:

$$TF(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

$$IDF(t) = log\_e\ (\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it})$$

[10] http://www.tfidf.com/

6.1.2 Cosine Similarity

Cosine Similarity is used to estimate the degree of similarity between the documents (Lahitani et al.; 2016). The similarity measure is used to calculate the similarity distance between the search query and the feature matrix prepared using SVD and Doc2Vec. In this research, we have calculated BoW(Bag of Words) similarity using BoW and SVD feature matrix and Semantic similarity using semantic message array and doc2Vec feature matrix. Thus, to increase the accuracy of the model, ensemble similarity is calculated by averaging the bow similarity and semantic similarity. Based on the similarity scores with the higher similarity between the search query and documents used for training purpose, the most similar musical instruments are evaluated and these similar instruments are recommended to users. The Cosine similarity is calculated using the formula below:

$$\cos\theta = \frac{\overrightarrow{A}.\overrightarrow{B}}{||\overrightarrow{A}||||\overrightarrow{B}||}$$

6.2 Results

The basic strategy followed in evaluating recommender system is getting the relevant recommendation based on the search query entered by the user that contains the sentiments of a user in the form of feelings, emotions, description, and brand. The results obtained using the recommendations were based on the human evaluation, how relevant they were from the users perspective. The results obtained using this recommendation system were specifically based on how detailed the users description was while finding the instrument. The recommendations obtained were satisfactory in a number of cases and also unsatisfactory in few cases. The experiments performed for evaluating the recommendation system are as follows.

To test how relevant the recommendation is, from the train data set, the description and brand of the musical instrument are entered as a search query using a chatbot. As we have referred our train dataset, we already have knowledge about what the expected output should be. Therefore, the expected output and the actual output from the chatbot is matched to check whether our system is successful in recommending items of users interest.

6.3 Experiment 1

Input Search Query: "Beginners and young learners alike will appreciate the quality found in this Yamaha C series classical guitar. This quality instrument delivers outstanding cost performance with exceptional playability and tone. The C40 is a full-size nylon-string guitar by Yamaha."

Expected Output:"Yamaha C40 Full Size Nylon-String Classical Guitar"

Actual Chatbot Output:

(Refer Figure 8)

Resultant similarity scores for the musical instruments that matched the search query generating the top 3 relevant recommendations are displayed in table 4.The individual similarities were calculated using the Doc2Vec model LSA model, known as semantic similarity and BoW similarity for all the music instruments and sorted with the highest similarity scores at the top. For the instrument " Yamaha C40A Nylon-String Full-Size Classical Ac..", the semantic similarity was found to be 0.703274 and BoW (Bag of Words)

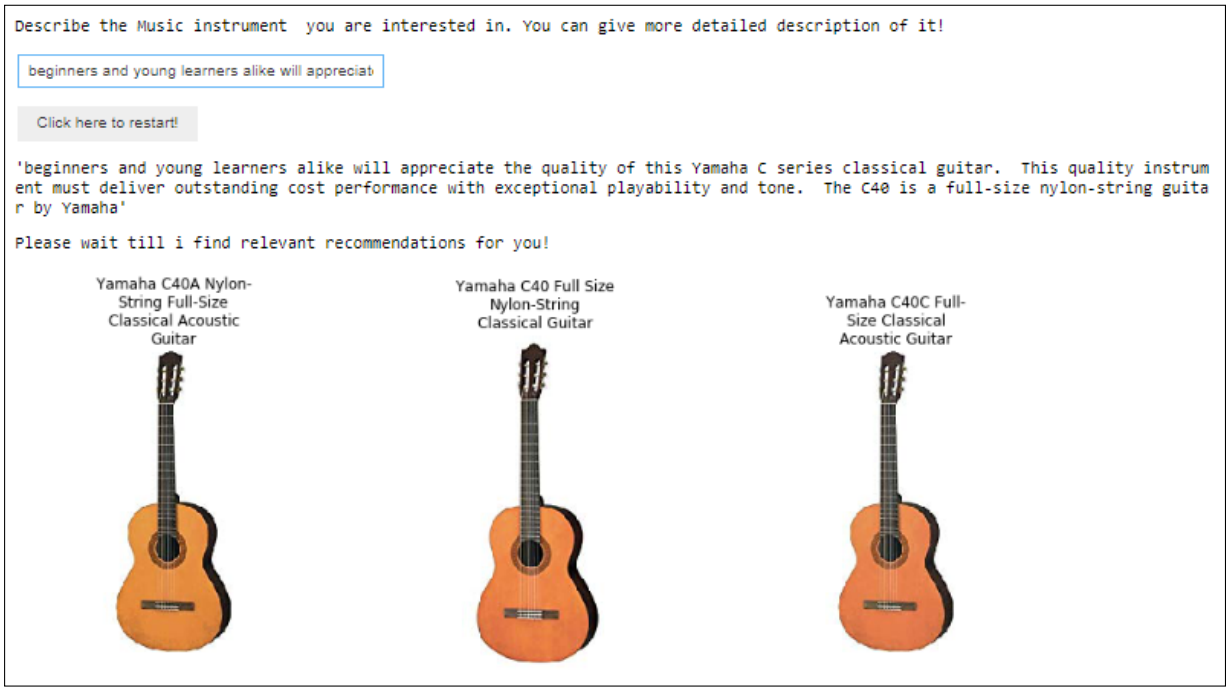

Figure 8: Output of Experiment 1

similarity as 0.947363. To get better results, these 2 similarities were averaged, generating the ensemble similarity and matched with the similarity measure generated for the search query entered through the chatbot. Similarly performed for the related instruments and filtered the top 3 relevant recommendations with the highest cosine similarity measure as shown in table 4. Thus the top 3 relevant recommendations are evaluated using cosine similarity.

Table 4: Cosine similarities of output

Title	**Semantic Similarity**	**BoW Similarity (Using SVD)**	**Ensemble Similarity (Average)**
Yamaha C40A Nylon-String Full-Size Classical Ac...	0.703274	0.947363	0.825319
Yamaha C40 Full Size Nylon-String Classical Guitar	0.725418	0.889326	0.807372
Yamaha C40C Full-Size Classical Acoustic Guitar	0.719851	0.859642	0.789747

6.4 Experiment 2

Input Search Query: "I need a ebonite (hard rubber) mouthpieces have beautifully proportional facing curves and tip openings by Yanagisawa"

Expected Output:"Yanagisawa Hard Rubber Alto Saxophone Mouthpiece 5"

Actual Chatbot Output:

(Refer Figure 9)

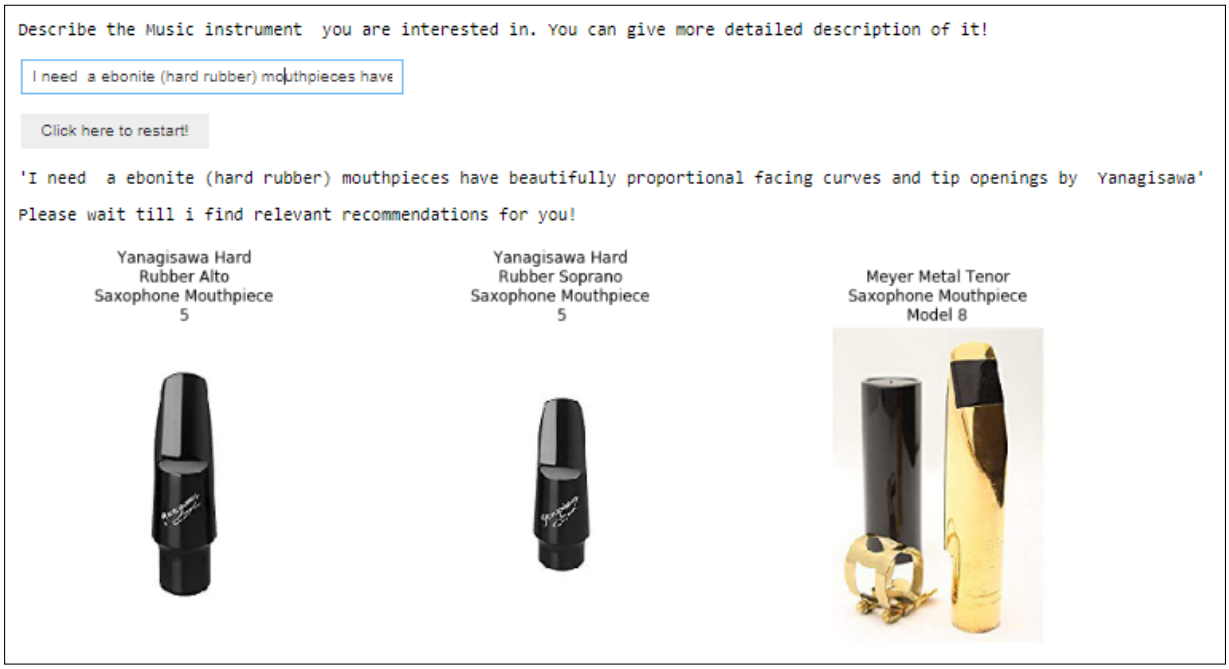

Figure 9: Output of Experiment 2

Resultant similarity scores for the musical instruments that matched the search query generating the top 3 relevant recommendations are displayed in table 5.The individual similarities were calculated using the Doc2Vec model LSA model, known as semantic similarity and BoW similarity for all the music instruments and sorted with the highest similarity scores at the top. For the instrument " Yanagisawa Hard Rubber Alto Saxophone Mouthpiece 5, the semantic similarity was found to be 0. 863261and BoW similarity as 0. 966511 and similarly for other instruments. To get better results, these 2 similarities were averaged, generating the ensemble similarity and matched with the similarity measure generated for the search query entered through the chatbot. Similarly performed for the related instruments and filtered the top 3 relevant recommendations with the highest cosine similarity measure as shown in table 5. Thus the top 3 relevant recommendations are evaluated using cosine similarity.

Table 5: Cosine similarities of output

Title	**Semantic Similarity}**	**BoW Similarity (Using SVD)**	**Ensemble Similarity (Average)**
Yanagisawa Hard Rubber Alto Saxophone Mouthpiece 5	0.863261	0.966511	0.914886
Yanagisawa Hard Rubber Soprano Saxophone Mouthp...	0.846524	0.957810	0.902167
Meyer Metal Tenor Saxophone Mouthpiece Model 8	0.794562	0.936607	0.865585

6.5 Experiment 3

As we can see from the above two experiments performed, we get relevant recommendations and are acceptable results. But in some cases, if the description in the search query and the document with which the model is trained doesnt match, we get irrelevant recommendations which is shown in below chatbot output.

Input Search Query: "I need a Gold plated metal mouthpiece."

Expected Output:"Something that is related to Gold plated metal mouthpiece"

Actual Chatbot Output:

(Refer Figure 10)

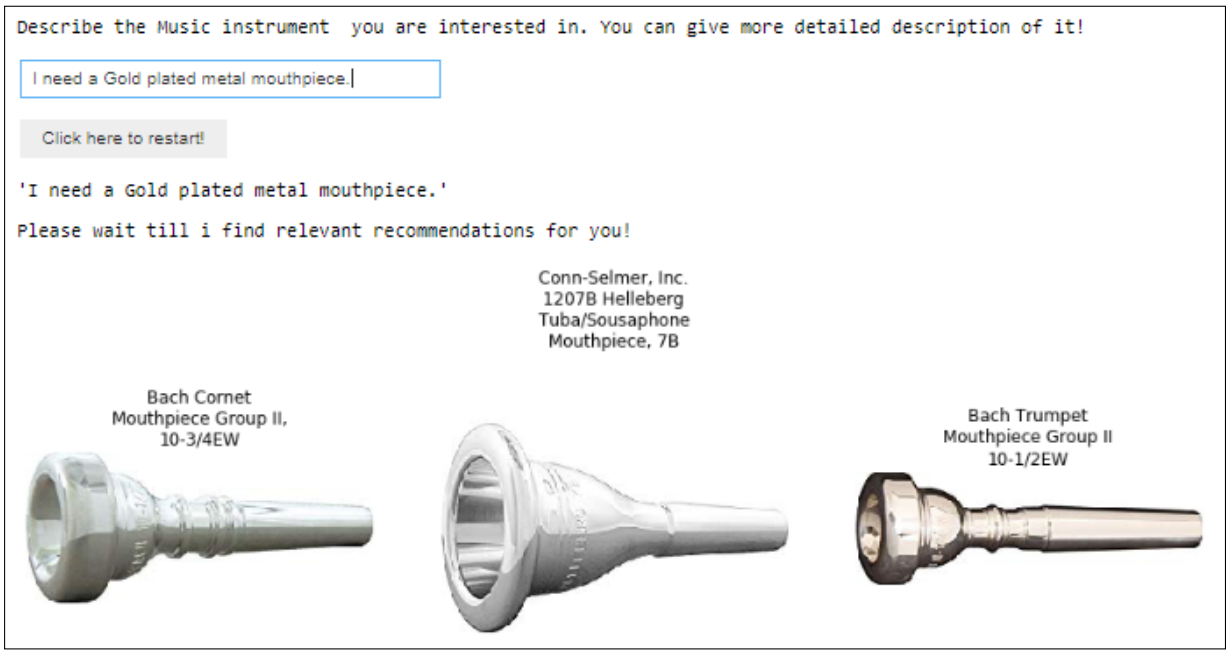

Figure 10: Output of Experiment 3

Resultant similarity scores for the musical instruments that matched the search query generating the top 3 relevant recommendations are displayed in table 6

Table 6: Cosine similarities of output

Title	Semantic Similarity	BoW Similarity (Using SVD)	Ensemble Similarity (Average)
Bach Cornet Mouthpiece Group II, 10-3/4EW	0.843399	0.938136	0.890767
Conn-Selmer, Inc. 1207B Helleberg Tuba/Sousapho...	0.818500	0.963005	0.890753
Bach Trumpet Mouthpiece Group II 10-1/2EW	0.833174	0.924696	0.878935

Thus, the above experiments performed shows how efficiently a recommendation system recommends the musical instruments to users based on the description, sentiments or the brand. The experiments 1 & 2 shows based on the search query, how the expected output and actual output matched as the model was trained with the actual descriptions and reviews. But considering the 3rd experiment, both the models were less accurate to recommend what exactly was required and thus instead of Gold plated it recommended the mouthpiece with Silver plated. Hence, there are certain limitations that needs be considered which is discussed in section 7 of this research.

7 Discussion

This research follows the use of Natural Language Processing approach to recommend the musical instrument on the basis of the user's description. This description is mainly in the form of the search query entered on a chatbot that contains a description of the instrument, sentiments of the user towards the instrument and the brand user is interested in. Thus, the above experiments performed shows that the system was successfully able to recommend the items related to the search query. Also, we can observe that, in experiment 3, then recommendations were not so accurate as compared to the first two. Though the models LSA and Doc2Vec were successfully able to recommend the items based on the data they were trained, we also need to consider different factors that will increase the accuracy of the recommendations and provide more relevant ones. The dataset used in this research had certain limitations. These limitations were mostly related to the details of the musical instrument. The data set lacked details like the Notes related to instruments, technical specifications of the instrument, description of color as a separate attribute of the dataset. If these limitations are overcome in future we would get better and more accurate recommendations. As we have performed a limited number of experiments, it is hard to decide how well our models work and give relevant recommendations. But to test the models and monitor the performance of how they truly work, this system should be deployed in A/B testing environment, and analyze whether the customer using this system, truly buys the items recommended.

8 Conclusion and Future Work

This research was performed using the dataset from Amazon to generate the recommendations of user's interest while buying a musical instrument. We have found that models: LSA (Latent Semantic Analysis) and Doc2Vec (Paragraph Vectoring) have performed very well in generating recommendations with the help of feature extraction from the training data and generating feature matrix. The experiments performed in the above section 6.2 shows the efficiency of the models in recommending the musical instruments. Thus, we can say that the choice of models and approach followed in building this system helped to gain more accurate results. Thus, the use of chatbot for answering the queries entered by user based on the description, sentiments, and brand of musical instrument, helped the system to be more interactive generating relevant results.

This type of system prepared for recommending the musical instruments can be generalized to recommend any type of product with a specific category for e.g. books, electric appliances, perfumes, clothing, shoes, etc. An important advantage of this type of system is, it would consider the sentiments of the user and ask for detail description, to give best

out of it. This type of system deployed in a live environment, in the near future, would generate business and be more competitive in today's world of e-commerce.

9 Acknowledgement

I would like to express my sincere gratitude towards my supervisor Dr. Anu Sahni, who always guided me and kept me motivated throughout my research work. It also helped me to gain a lot of knowledge while seeking her guidance in this research. Also, I would like to thank my family who constantly supported me throughout this research work and at last, my friends who always helped me in solving the difficulties I faced.

References

Alspector, J., Kolcz, A. and Karunanithi, N. (1998). Comparing feature-based and clique-based user models for movie selection, *Proceedings of the third ACM conference on Digital libraries*, ACM, pp. 11–18.

Armstrong, R., Freitag, D., Joachims, T. and Mitchell, T. (2003). Webwatcher: A learning apprentice for the world wide web.

Bobadilla, J., Ortega, F., Hernando, A. and GutiéRrez, A. (2013). Recommender systems survey, *Know.-Based Syst.* **46**: 109–132.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. R. H. and Wirth, R. B. (2000). Crisp-dm 1.0: Step-by-step data mining guide.

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B. and Sampath, D. (2010). The youtube video recommendation system, *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, ACM, New York, NY, USA, pp. 293–296.

Gomez-Uribe, C. A. and Hunt, N. (2016). The netflix recommender system: Algorithms, business value, and innovation, *ACM Transactions on Management Information Systems (TMIS)* **6**(4): 13.

Halko, N., Martinsson, P. G. and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, *SIAM Rev.* **53**(2): 217–288.

Han Lau, J. and Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation, pp. 78–86.

Hanani, U., Shapira, B. and Shoval, P. (2001). Information filtering: Overview of issues, research and systems, *User Model. User-Adapt. Interact.* **11**: 203–259.

Herlocker, J. L., Konstan, J. A., Borchers, A. and Riedl, J. (1999). An algorithmic framework for performing collaborative filtering, *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999*, Association for Computing Machinery, Inc, pp. 230–237.

Hutto, C. and Gilbert, E. (2015). Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Isinkaye, F., Folajimi, Y. and Ojokoh, B. (2015). Recommendation systems: Principles, methods and evaluation, *Egyptian Informatics Journal* **16**(3): 261–273.

Kucherbaev, P., Psyllidis, A. and Bozzon, A. (2017). Chatbots as conversational recommender systems in urban contexts, *Proceedings of the International Workshop on Recommender Systems for Citizens*, CitRec '17, ACM, New York, NY, USA, pp. 6:1–6:2.

Lahitani, A. R., Permanasari, A. E. and Setiawan, N. A. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment, *2016 4th International Conference on Cyber and IT Service Management*, pp. 1–6.

Landauer, T. K., Foltz, P. W. and Laham, D. (1998). An introduction to latent semantic analysis, *Discourse processes* **25**(2-3): 259–284.

Lang, K. (1995). Newsweeder: Learning to filter netnews, *in* A. Prieditis and S. Russell (eds), *Machine Learning Proceedings 1995*, Morgan Kaufmann, San Francisco (CA), pp. 331 – 339.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents, *International conference on machine learning*, pp. 1188–1196.

Linden, G., Smith, B. and York, J. (2003). Amazon.com recommendations: item-to-item collaborative filtering, *IEEE Internet Computing* **7**(1): 76–80.

Lu, J. (2004). A personalized e-learning material recommender system, *International Conference on Information Technology and Applications*, Macquarie Scientific Publishing.

Lund, J. and Ng, Y. (2018). Movie recommendations using the deep learning approach, *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 47–54.

Ma, H., Wang, X., Hou, J. and Lu, Y. (2017). Course recommendation based on semantic similarity analysis, *2017 3rd IEEE International Conference on Control Science and Systems Engineering (ICCSSE)*, pp. 638–641.

Mooney, R. J. and Roy, L. (2000). Content-based book recommending using learning for text categorization, *Proceedings of the fifth ACM conference on Digital libraries*, ACM, pp. 195–204.

Nath Nandi, R., Arefin Zaman, M. M., Al Muntasir, T., Hosain Sumit, S., Sourov, T. and Jamil-Ur Rahman, M. (2018). Bangla news recommendation using doc2vec, *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–5.

Pan, C. and Li, W. (2010). Research paper recommendation with topic analysis, *2010 International Conference On Computer Design and Applications*, Vol. 4, pp. V4–264–V4–268.

Pazzani, M. J. and Billsus, D. (2007). *Content-Based Recommendation Systems*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 325–341.

Pazzani, M., Muramatsu, J. and Billsus, D. (1998). Syskill webert: Identifying interesting web sites, *Proceedings of the National Conference on Artificial Intelligence* **1**.

Pu, P., Chen, L. and Hu, R. (2011). A user-centric evaluation framework for recommender systems, *Proceedings of the fifth ACM conference on Recommender systems*, ACM, pp. 157–164.

Schafer, J. B., Frankowski, D., Herlocker, J. and Sen, S. (2007). Collaborative filtering recommender systems, *The adaptive web*, Springer, pp. 291–324.

Schafer, J. B., Konstan, J. and Riedl, J. (1999a). Recommender systems in e-commerce, *Proceedings of the 1st ACM conference on Electronic commerce*, ACM, pp. 158–166.

Schafer, J. B., Konstan, J. and Riedl, J. (1999b). Recommender systems in e-commerce, *Proceedings of the 1st ACM Conference on Electronic Commerce*, EC '99, ACM, New York, NY, USA, pp. 158–166.

Tan, H., Guo, J. and Li, Y. (2008). E-learning recommendation system, *2008 International Conference on Computer Science and Software Engineering*, Vol. 5, pp. 430–433.