IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Drug Recommendation System on Imbalanced Medical Data Using SMOTE

**YELIKE SAI VENKATA VAMSI KRISHNA**[#1]

[#1] B.Tech Student,

Department of Computer Science & Engineering,
Sikkim Manipal Institute of Technology, Majitar, Rangpo Sikkim State -737136.

vamsiyelike2000@gmail.com [#1]

**ABSTRACT**

Nowadays there is a lot of imbalanced data present on the web in several areas like medical, business, insurance, banks, research centers, and so on. All these imbalanced data are creating a lot of gaps for the researchers to gain any new information based on the primitive data. Since the corona virus emerged, it has been increasingly difficult to get legitimate therapeutic resources, such as the scarcity of specialists and healthcare professionals, appropriate equipment and medications, etc. There are many deaths as a result of the medical profession as a whole being in distress. People started taking medications independently without the proper consultation because they weren't readily available, which made their health conditions worse than usual. Recently, machine learning has proven useful in a variety of applications, and creative work for automation is on the rise. The goal of this research is to present a medicine recommendation system that significantly decreases the need for specialists. Because they are doctor-friendly, emerging technologies like machine learning, deep learning, and data mining can help us investigate the history of medicine and lower medical errors. In order to identify the top few medications for a condition, this research suggests a drug recommendation system that uses word2vec sentiment analysis on patient review data.

**Index Terms:**

Word2Vec Model, Sentiment Analysis, Healthcare Professional, Recommendation System, Imbalanced Data.

## 1. INTRODUCTION

The internet is one of the places where people are most worried and look up information on various topics. Nearly 60% of adults, according to the Pew Internet and American Life Project, are looking for adequate health information online, with 35% of respondents focusing solely on online disease diagnosis. Since several studies reveal that many people lose their lives as a result of medical mistakes made by healthcare professionals who prescribe medications based on their experiences. They frequently make blunders because the majority of their experience is limited. The world is experiencing a doctor shortage due

to the exponential increase in corona virus cases, especially in rural areas where there are fewer experts than in urban areas.

A doctor must complete their education between six and twelve years. Therefore, it is impossible to increase the number of doctors in a short period of time. Clinical errors occur often today. Every year, medication errors have an impact on over 200 000 people in China and 100,000 people in the USA. Since specialists only have a limited amount of information, they frequently prescribe the wrong medication (about 40% of the time). This project offers doctors a medication recommendation system they can use when writing prescriptions. When delivering patients their medications, this is also beneficial for the pharmacists. The UCI ML repository, which is online, provided the data for the research.

This dataset has six attributes: the name of the drug used (text), the patient's review (text), the patient's condition (text), the useful count (numerical), the date (date) of the review entry, and a 10-star patient rating (numerical) indicating overall patient satisfaction. There are altogether 215063 instances in it. The dataset's "review" column underwent several text pre-processing techniques like stemming, lemmatization, stop-word removal, etc. The numerical characteristics were cleaned and adjusted as necessary. To determine whether a sentiment is positive or negative, use the "user-rating" feature. Combining multiple machine learning classifiers, such as Linear Support Vector Classifier, Decision Tree, Random Forest, Light, and Bag of Words (BOW), Term Frequency-Inverse Document Frequency (TF-IDF), Word2vec, and N-gram models are applied.

## 2. LITERATURE SURVEY

In this section we will mainly discuss about the background work that is carried out in order to prove the performance of our proposed Method. Literature survey is the most important step in software development process. For any software or application development, this step plays a very crucial role by determining the several factors like time, money, effort, lines of code and company strength. Once all these several factors are satisfied, then we need to determine which operating system and language used for developing the application. Once the programmers start building the application, they will first observe what are the pre-defined inventions that are done on same concept and then they will try to design the task in some innovated manner.

### MOTIVATION

A well known author Leilei Sun, et.al, proposed a paper" Optimal treatment regimens a patient's Electronic Medical Record (EMR)," To get the optimal therapy recommendation for patients, looked through a huge number of treatment records. The plan was to estimate the similarity between treatment data using a powerful semantic clustering technique. The author also developed a framework to evaluate the suitability of the recommended course of treatment. According to the demographics and medical difficulties of new patients, this framework can recommend the optimal treatment regimens. a patient's Electronic Medical Record (EMR) that has been collected for testing from several clinics. The outcome demonstrates how this paradigm raises the cure rate.

A well known author, Xiaohong Jiang ,et.al, proposed a work and explained as : On the basis of treatment data, investigated three different algorithms: the decision tree algorithm, the support vector machine (SVM), and the back propagation neural network. Model exactness, model proficiency, and model variety were the three distinct criteria that were used to select SVM for the medicine suggestion module. Additionally, a system for checking for errors was suggested to guarantee the accuracy of the administration, analysis, and data.

A well known author, Mohammad Mehedi Hassan et al., [7] proposed a framework as " CADRE can recommend medications with top-N related prescriptions based on the adverse effects

experienced by patients". This proposed framework's basic foundation was made up of collaborative filtering procedures, in which the drugs are initially grouped into clusters based on the functional description data. However, the model is changed to a cloud-aided technique employing tensor decomposition for improving the quality of experience of medication suggestion after taking into account its shortcomings, such as computationally expensive, cold start, and information sparsity.

Jiugang Li et al. [8] developed a hashtag recommender system using the skip-gram model and convolutional neural networks (CNN) to learn semantic phrase vectors, taking into account the importance of hashtags in sentiment analysis. These vectors employ LSTM RNN to classify hashtags based on the features. Results show that this model outperforms more widely used models like SVM and Standard RNN. This investigation is based on the fact that it was subjected to standard AI approaches like SVM and collaborative filtering; the semantic features are lost, which has a significant impact on obtaining a reasonable expectation.

The Panacea [9] framework for discovering drug-drug and drug-interaction interactions is based on the Galen OWL method and makes use of a comprehensive knowledge base and standardised medical words that are both modelled as rules. They made use of the SKOS lexicon, an ontology and reasoning engine, as well as a rules-based and medical approach to reasoning. The findings indicate that Panacea is a promising approach, although it still requires additional work.

Ontology-based methodology is used by SemMed [10], a medical recommendation engine built on Semantic Web Technologies. It includes an ontology manager, rules manager, inference engine, and support database. Rules were created using the fundamental classes "Diseases," "Medicines," and "Allergies."

The IRS-T2D [12] drug recommender system was created primarily to individualise patient care for people with type 2 diabetes mellitus. The approach incorporates rule-based decision making, ontologies, and semantic web technologies while taking into account particular patient data, such as the unique HbA1c target.

## 3. EXISTING SYSTEM & ITS LIMITATIONS

In the existing system there was no proper method to identify the patient symptoms and give medicine accordingly to those corresponding patients. All the existing systems try to take medicines based on physical visit to the hospital. There is no appropriate technique which can recommend the drugs based on the symptoms what is identified in the patient. The existing system follows manual approach and hence the following are the limitations of the existing system.

1. More Time Delay in finding the symptoms and problems from the patient.
2. There is a huge delay in finding the disease.
3. All the existing approaches are manual approach and hence it is very complex task for the medical person to collect the details from the patients.
4. There is no recommendation system in the existing system.

## 4. PROPOSED SYSTEM & ITS ADVANTAGES

In the proposed system, we try to construct an application which can give medicine recommendation that significantly decreases the need for specialists. In general all the primitive methods try to use manual approach to identify the disease and provide drugs to that particular disease. In this current system we try to construct doctor-friendly, emerging technologies like machine learning, deep learning, and data mining to lower the medical errors. In order to identify the top few medications for a condition, this research suggests a drug recommendation system that uses word2vec sentiment analysis on patient review dataThe following are the advantages of the proposed system. They are as follows:

1) By using Word2Vec sentiment analysis model on patients, we can able to recommend the drugs accurately for the corresponding patient.
2) In this proposed work, we try to classify the unbalanced data accurately.
3) The proposed model is very efficient in recommending the drugs.
4) In this proposed work we can construct user friendly for drug recommendation.

## 5. PROPOSED MODELS

In this proposed system we try to use SMOTE and Word2Vec model for recommendation of drugs for the patients based on the symptoms collected from that patient.

**Word2Vec Model**

This is one of the group models which is used for constructing the word embeddings. This model is used for pre-processing the text data into a shallow and two-layered manner to train the ML or DL models. This is mainly used for embedding the words with meanings, and those which are having the same meaning are clubbed into one group, and they are sent as input for the Ml or Deep Learning Algorithms.

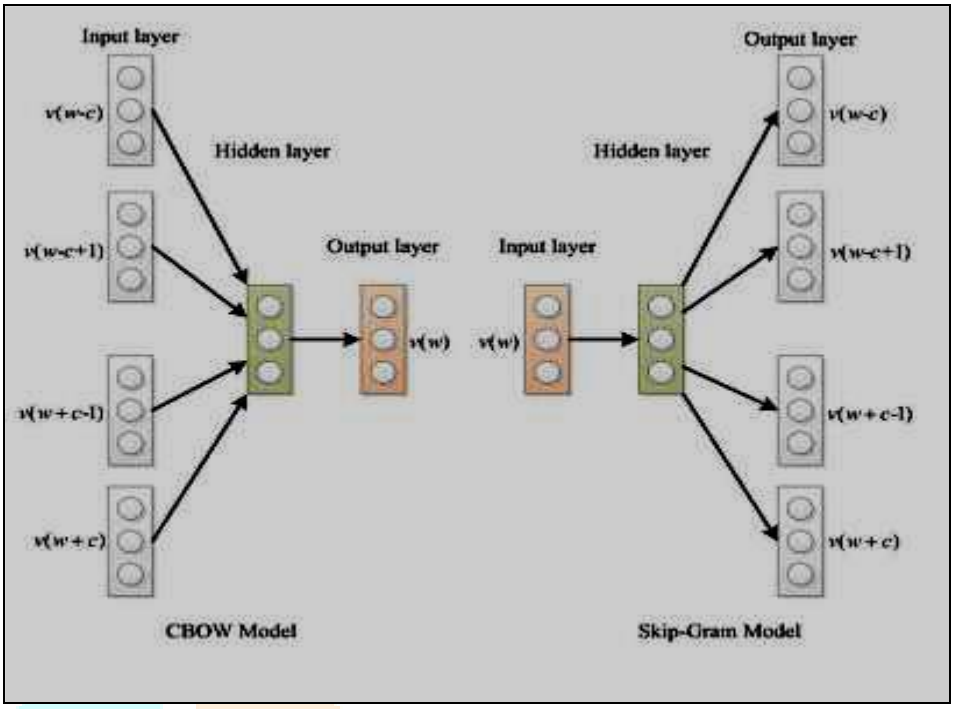



**Figure 1. Represents the Architecture of Word2Vec Word Embedding Model**

In general the word2vec model is having more efficiency in word embedding compared with several other models. Normally when we try to load any dataset, we will first extract the main features which are present in that dataset. Once the main features are extracted now we try to find out the words which are to be embedded. This word embedding is applied in order to form uniqueness. Those which are repeated multiple times.

## SMOTE (Synthetic Minority Oversampling Technique)

Creating prediction models for classification datasets with a significant class imbalance is known as imbalanced classification. Working with imbalanced datasets is a difficult because most machine learning approaches overlook the minority class, which results in poor performance even though this performance is often the most crucial. Oversampling the minority class is one way to deal with unbalanced datasets. Duplicating examples from the minority class is the simplest method, but these examples don't provide any new insight into the model. Instead, fresh examples can be created by synthesising the old ones. The Synthetic Minority Oversampling Technique, or SMOTE for short, is a type of data augmentation for the minority class.

In general for drugs to be recommended for the patients it is very difficult for the end users to recommend the drugs directly for the patients, hence we need to have a clear insight about the symptoms and then check which data is exactly synthesising with the correct drug and hence based on this the drug recommendation is done. SMOTE identifies the k closest minority class neighbours of a minority class instance an after choosing it at random. Next, a line segment in the feature space is formed by joining a and b

to produce the synthetic instance by randomly selecting one of the k nearest neighbours, b. The two selected examples, a and b, are convexly combined to create the synthetic instances.

SMOTE creates new minority instances by combining minority instances that already exist. For the mi nority class, it creates virtual training records using linear interpolation. For each example in the minority clas s, one or more of the knearest neighbours are randomly chosen to serve as these synthetic training records. Fol lowing the oversampling procedure, the data is rebuilt and can be subjected to several categorization models.

**Step 1:** Using the minority class set A as a starting point, the k-nearest neighbours of each $x in A$ are determined by determining the Euclidean distance between each sample in set A and $x.

**Step 2:** The uneven proportion is used to determine the sample rate N. N samples (i.e. x1, x2,...xn) are randomly chosen from each $x in A$'s k-nearest neighbours, and they combine to form the set $A 1$.

**Step 3:** To create a new example for each case $x k in A 1$ (k=1, 2, 3...N), apply the following formula: Where rand(0, 1) denotes a random number between 0 and 1, $x' = x + rand(0, 1) * mid x - x k mid$.

## 6. IMPLEMENTATION PHASE

Implementation is the stage where the theoretical design is converted into programmatically manner. In this stage we will divide the application into a number of modules and then coded for deployment. The front end of the application takes python and back end we take drug dataset from google. The application is divided mainly into following 6 modules. They are as follows:

1. Import Necessary Libraries
2. Load Dataset Module
3. Data Cleaning and Visualization
4. Feature Extraction
5. Train and Test Using SMOTE
6. Modeling and Evaluation

### 1. IMPORT NECESSARY LIBRARIES

In this module initially we need to import all the necessary libraries which are required for building the model. Here we try to use all the libraries which are used to convert the data into meaningful manner. Here the data is divided into numerical values which are easily identified by the system, hence we try to import numpy module and for plotting the data in graphs and charts we used matplot library.

### 2. LOAD DATASET MODULE

In this module the we try to load the dataset which is downloaded or collected from Kaggle repository. Here we store the drug related dataset and based on the information present in the dataset, we can able to extract the necessary information which is required for suggesting the drugs for the patients.

### 3. DATA CLEANING & VISUALIZATION MODULE

Here in this section we try to pre-process the input dataset and find out if there are any missing values or in-complete data present in the dataset. If there is any such data present in the dataset, the application will ignore those values and load only valid rows which have all the valid inputs.

### 4. FEATURE EXTRACTION

In this step the main features which are required for drug recommendation is first identified and all the information is processed with our model.Once the main features are extracted now we can able to find out the relation between each and every drug corresponding to the symptoms of that patients.

### 5. TEST & TRAIN USING SMOTE

In this module we need to divide the data into two categories: One is test and another is train based on the number of attributes. Here we use SMOTE to categorize the data based on primitive information.

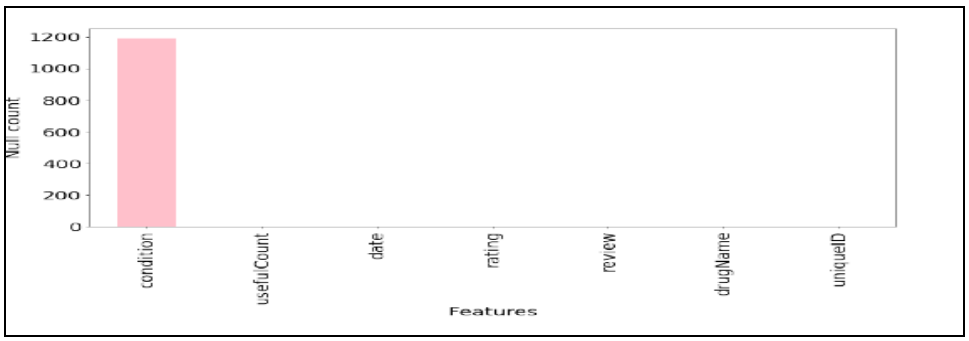
### 6. MODEL EVALUATION

In this module we can able to extract the models performance based on the time and efficiency compared with several other models. Hence this module is main module to get the evaluation result.

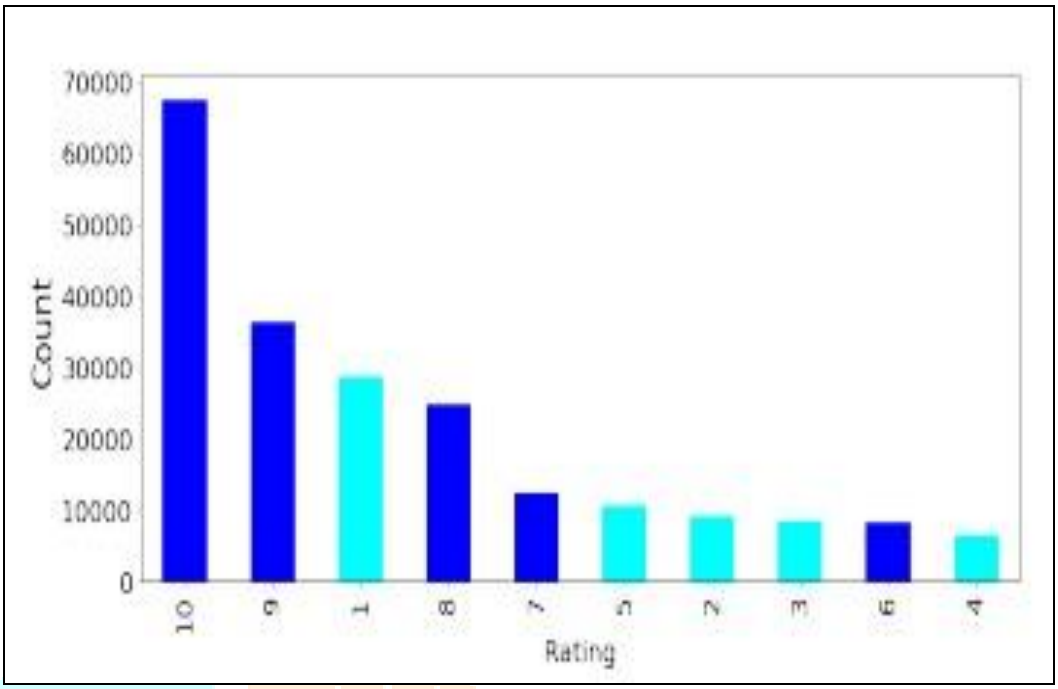
## 7. EXPERIMENTAL REPORTS

In this application we try to use Python as coding platform and take word2Vec and SMOTE model as training model to drug recommendation.

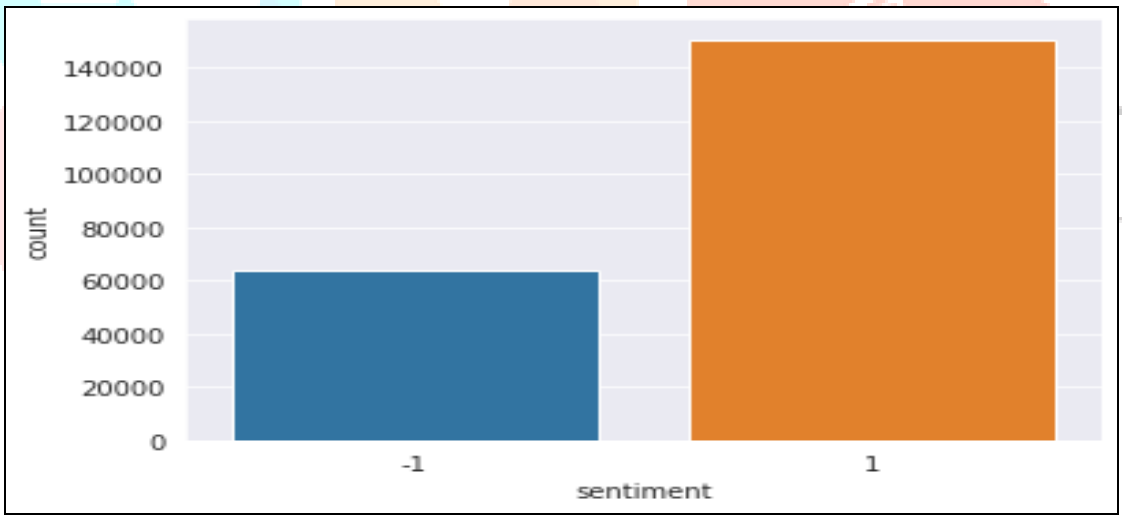
## Data Cleaning



## Count plot for Rating




## Distribution Chart




From the above window we can clearly identify that there are two possibilities in the above graph such as positive sentiment or negative sentiment and for positive sentiment we can see value as "1" and for negative sentiment,we can see the same value as "-1".

## 8. CONCLUSION

This project was intended to produce state-of-the-art results in forecasting medicines and suggesting them based on sentimental analysis of user evaluations and valuable count provided by them. It was motivated and inspired by the most recent trends and research in recommender systems. A multi-dimensional space of word vectors that are semantically related to one another was created using the word2vec model, one of the

various vectorization methods that are accessible. Prior to using the vectorization procedure, all required text pretreatment techniques were applied to the dataset's "review" feature. With just the train data in between, various Synthetic Minority Over Sampling Technique variations were tested, along with hyperparameter modification for the "sampling strategy". To compare their accuracy, later machine learning algorithms were put into use. Random Forest Classifier had a 67% accuracy rate compared to Decision Tree Classifier's 50%. The Light Gradient Boosting classifier provided a 70% accuracy out of all the techniques used. The normalized useable count was multiplied by the generated predictions to obtain the final mean score for each drug. The score was then divided by the number of medicines per condition.

## 9. REFERENCES

[1] T. N. Tekade and M. Emmanuel, "Probabilistic aspect mining approach for understanding and evaluation of drug reviews," in 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, pp. 1471–1476.

[2] GalenOWL: Ontology-based drug recommendations discovery. Doulaverakis, C., Nikolaidis, G., Kleontas, et al 13 J Biomed Semant (2012). https://doi.org/10.1186/2041-1480-3-14

[3] Hui Xiong, Yanming Xie, Chuanren Liu, Leilei Sun, and Chonghui Guo. 2016. Development and Recommendation of Automatic Treatment Regimes Based on Data. The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16) s

[4] "Sentiment Analysis of Multilingual Twitter Data Using Natural Language Processing," by V. Goel, A. K. Gupta, and N. Kumar. 2018 8th

[5] Drug-recommendation system for patients with infectious disorders. Shimada K, Takada H, Mitsuyama S, et al. 2005;2005:1112; AMIA Annu Symp Proc.

[6] Y. Bao and X. Jiang, "An intelligent medicine recommender system framework," 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), Hefei, 2016, pp. 1383–1388.

[7] Yin Zhang, Dafang Zhang, Mohammad Hassan, Atif Alamri, and Limei Peng. (2014). For online pharmacies, CADRE stands for Cloud-Assisted Drug Recommendation Service. Mobile Applications and Networks. 20. 348-355. 10.1007/s11036-014-0537-4.

[8] Tweet modelling with LSTM recurrent neural networks for hashtag recommendation; 2016 International Joint Conference on Neural Networks (IJCNN); Vancouver, BC; pp. 1570–1577; doi: 10.1109/IJCNN.2016.7727385.

[9] Kleontas, A., Nikolaidis, G., and Kompatsiaris, I. Doulaverakis, C., et al. Panacea is a framework for discovering medicine recommendations with semantic support. 5, 13, and Journal of Biomedical Semantics

[10] Rodrguez, A., Jimenez, E., Fernandez, J., Eccius, M., Gómez, J. M., Alor-Hernandez, G., Posada-Gomez, R., and Laufer, 2009. Semantic Web for Medical Recommendation Systems, or SemMed. The first global conference on intensive applications and services was held in 2009.

[11] according to Chen, R.-C., Chiu, J. Y., and Batj, C. T. An example of an anti-diabetic medication is provided in The recommendation of Medicines based on Multiple Criteria Decision Making and Domain Ontology, 27–32.

[12] N. Mahmoud and H. Elbeh. IRS-T2D. Individualize Type 2 Diabetes Medication Recommendation System Based on Ontology and SWRL. 10th International Conference on Informatics and Systems Proceedings: INFOS '16. ACM

[13] Y.-H. Huang, C.-T. Bau, R.-C. Chen, S.-M. Chen, and others 2012. a method for selecting diabetes medication based on SWRL and domain ontology. Expert Systems with Applications 39, 4 (2009, pp. 3995–406).

[14] Time-aware Twitter-based drug recommender system, T-Recs, by A. A. Hamed, R. Roose, M. Branicki, and A. Rubin. International Conference on Social Network Analysis and Mining, 2012 IEEE/ACM.