Are story outlines confined within seven basic plots?

MSc Research Project
Data Analytics

Amith Anand Sethu
Student ID: 17163650

School of Computing
National College of Ireland

Supervisor: Dr Anu Sahni

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Amith Anand Sethu
Student ID:	17163650
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr Anu Sahni
Submission Due Date:	12/08/2019
Project Title:	Are story outlines confined within seven basic plots?
Word Count:	5984
Page Count:	18

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:	9th 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):	

Are story outlines confined within seven basic plots?

Amith Anand Sethu
17163650

Abstract

This research will discuss on a unique method to model story plot data by analyzing the pattern of events in the text content and then clustering the story plots into seven categories that would imply a story outline. Processing the story outlines and categorizing stories based on pattern of events in stories will aid in improving recommender systems. The idea behind this type of a model is to cluster data based on context from the text data rather than just topic words present in the data. In this research a FrameNet classifier is modeled to classify the sentences into frames. the frame classifier model was built using a Linear Kernal SVM. The model showed a 67% accuracy predicting the frames. The sequence of frames are transformed into numeric representations and then clustered using k means clustering algorithm to fit within seven clusters. The model was evaluated using an elbow test and silhouette analysis to verify the number of clusters chosen which showed a good result of 6 as optimal number of clusters.

1 Introduction

Every story is unique its own theme, characters, place and narration. What makes a story a success is a variety of parameters. It is a mixture of a variety of characters, places, time and the conflicting point(Morris; 2008). The combination of all the parameters sets the feel around the movie. A story narrated in a specific structure is still an area quite unexplored in machine learning. Out of all the parameters that are affecting, the narration parameter is the most effective one. As it can be been observed, in lot of movies wherein technological advancements would have been made or presence of great characters in the story but lacks a good narration seems to have failed most of the time.

People can distinguish between a good narration and a bad narration. A person understands the pattern that follows in stories and so the person can even predict what happens next based on the narration style. Ever since Alan Turing came up with Turing machine, mankind's approach has always been automating everything in the world to speed up tasks or automate tasks which is not humanly possible. In the current age of information age, movies are classified using machine learning models and most of these models are classifying based on movie categories or movie reviews. But the existing classifications(Ertugrul and Karagoz; 2018) such as mystery, thriller, drama etc does not provide any sort of story outlines. These classifications are basically done by the means of the words in the text. Using a bag of words model or a neural network model that identifies the word vector values to classify the stories still does not consider the events taking place in the story. A well narrated story always has a pattern of events which

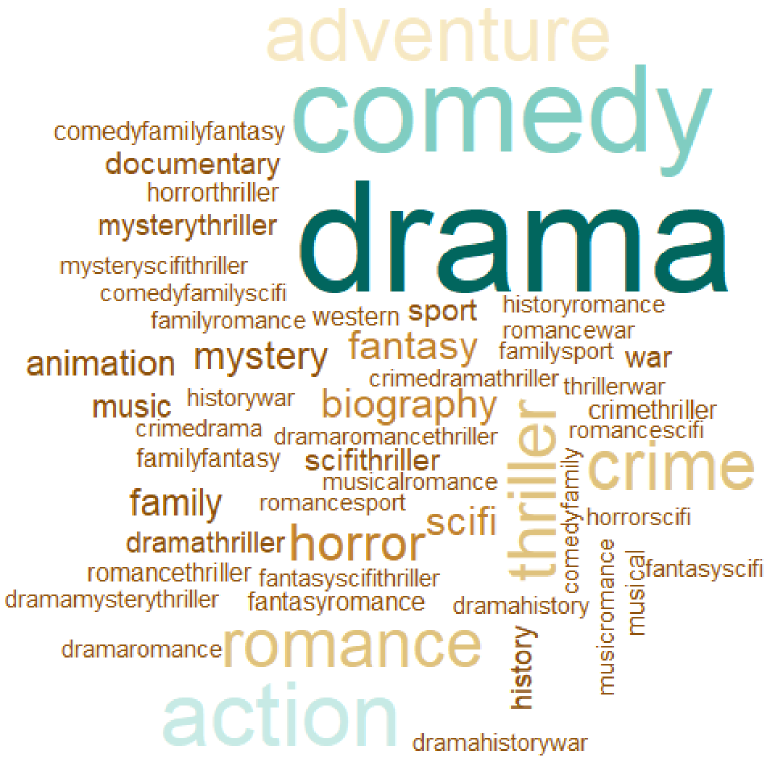

Figure 1: Genres

evoke the curiosity in the audience.

The motive behind this study is to discuss the importance of a machine learning model that classifies the story plots based on the event pattern in the stories and bring light to the idea of classifying text data(Kar et al.; 2018) based on events in the text data. Story summaries are a sequence of events in stories. This is the main concern in this research, the data that needs to be analyzed in story summaries has to be be primarily event data. The sequence of events would help shedding light on a pattern in story summary data and that the pattern in data would imply the existence of structure within summary data. Hence all story summaries will fall under a limited fixed number of unique clusters.

The research is focused on the idea for identifying a structural pattern in raw text data of story summaries by finding a pattern of event features extracted by using the frame concept introduced in FrameNet(Baker et al.; 1998) lexical data base. The research explores the features of FrameNet to build a classifier model that would classify the sentences in to frames. The identified frames would semantically make sense of the context in the data. The sequence of frames is then clustered using an unsupervised algorithm(KMeans clustering)(Krishna et al.; 2012) to cluster the data to identify the natural pattern present in the story plots. The number of clusters that was expected to be present in the data for the study is 7 which was based on the theory stated in the book 'The seven basic plots'(Booker; 2004) that all stories falls under just seven categories. Categories that has specific story outlines.

2 Related Work

Past researches have shown a variety of methods for data classification of movie data. One case would be a machine learning models using Multinomial probabilistic approach to predict the genre of the movie. Wherein a likelihood of a movie genre is established and using Bayesian Probabilistic reasoning(Makita and Lenskiy; 2016) the genre would be predicted. The feature used for such models are the words in the data. The word probability defines the genre prediction. These models are beneficial in applications of movie recommendation systems. The idea behind such models are based on the existence of topic words in data. Processing the data to remove all noise data and keeping only the topic words helps in clustering the data into specific clusters based on word usage in a movie category.

2.1 Plot classification

Basing the former research(Makita and Lenskiy; 2016) as an inspiration further research was done using the Bayesian Probablistic model. The research added two methods other than the existing Naive Bayes Classifier model to improve the model. Theses methods were added in concern of the context sense of the text data. The second method was based on word2vec and XGBoost(Hoang; 2018). The word2vec model generates the vector value of words based on the surrounding words and the XGBoost would give a probability distribution of movie genres. The final method used for the research was a recurring neural network model(RNN)(Hoang; 2018). The RNN model learns word embedding using its internal state to identify a pattern of words for a document.

Another research that employed a Bi-directionla LSTM model(Ertugrul and Karagoz; 2018) followed a unique idea of labelling every sentences in the plot with its associated genre. These labelled sentences was then used to train a model for learning the posterior probability of genre for its corresponding sentence. Using the probabilities, a majority voting would be then employed within a story by considering all sentence labels together. This method again uses word representations for classifying the documents into specific genres.

Natural language processing is an ongoing research field where a variety of techniques can be discovered to make sense of data. Existing models such as the word2vec and doc2vec(Le and Mikolov; 2014) were discovered for applying these models to form a vector representational model of words and documents. This way of representations aids in visualizing the data and transforming the data into a structural format that can be easily clustered or classified or understood. Researches have been conducted using word embedding(Kar et al.; 2018) models for classification basing the same idea of word usage in specific topics. Existing movie classes or genre such as mystery, drama, horror is more generic which does not imply the outline of a story. These classes rather implies the emotional label discussed in the story summaries. Emotion does not have any specific structure. A romantic movie does not imply any specific structure in the movie. The story could be about a protagonist chasing the love or it could be a story about a series of events happening in couple relationship.

2.2 Story plot structure

A good narrative is vital to a story. And it is not only in stories that narration is vital. The importance of narration extends to public speaking as well. Human minds have affinity towards an orderly set of events, that glues their minds to the source of the information whether it is a movie, story or even a speaker. What makes a public speaker good is his or her ability to put the series of words in an orderly structure which everyone would want to understand. The key to unlocking such an ability is placing the events in an appropriate chronological order(Hoeken and van Vliet; 2000) evoking the emotions in human mind. Structuring the content would also aid in generative writing(Motro; 2015) and eventually engage the readers. In a book 'Talk like ted'(Gallo; 2016) describes the art of speaking as means to tell to story by connecting with the listeners. Connecting the right set of sentences i.e not just connecting right set of words, the sentence sequences aids in building an emotion in the story and this adds life into the story or the speech.

A popular and well known method for structuring story content is the three act structure(Morris; 2008). Three act structure is a very old concept that has been followed by most authors for writing the stories. A study conducted on script writing states that when a story is confined within a structure it inhibits the organic development(Morris; 2008) of the story in human minds but a structure can rather be used in the development phase as mentioned earlier structure aids in generative writing.

2.2.1 Three Act Structure

This is a standard format for representing the structure of a story plot which divides a plot into 3 sections. The following are the 3 sections:

- Exposition
- Confrontation
- Resolution
  (Motro; 2015)

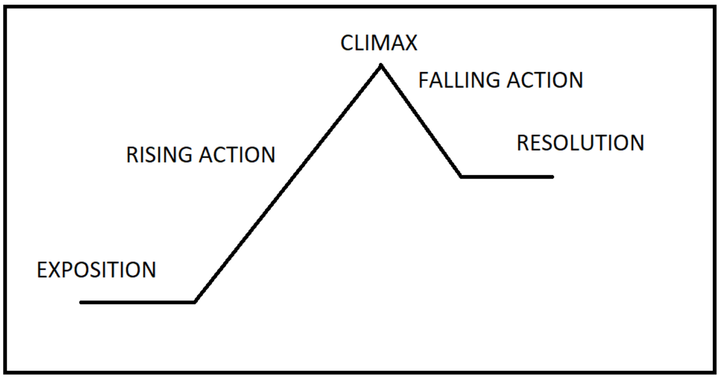

Figure 2: Three Act Structure(Abernathy and Rouse; 2014)

The exposition is the phase where the conflict in the plot is introduced and then the next section is accompanied with a series of events which is the rising action that builds curiosity in the audience. Following both these sections, the final section is the resolution section where the conflict comes to an end. This is a model many authors follow when writing scripts.

2.3 Context extraction

Existing researches have shown the importance of context extraction as an improvement over the topic word extraction. Research conducted on context extraction for recommender systems(Lahlou et al.; 2013) have shown satisfactory results without stemming the words and keeping them in the raw form. Such methods are employed to stress on the factor that it is not just topic words that matters but the context formed by combining a group of words does matter. Most text classification models are programmed to classify based on the word usage in the text content. And so basing this idea models have been built to generate word embedding(Liu et al.; 2018) of the document to classify documents based on the words and the surround words.

Further researches have focused on building models based on task oriented word embedding. Distinguishing words that have similar vector representations but different sentiment polarities(Yu et al.; 2017). Another study indicated the lack of ability of such task oriented methods to address the realistic tasks of the words in the content. The study came up with an alternate method of dynamically modelling the features for a task and then generating the vector representations(Liu et al.; 2018).

2.3.1 Semantic role labelling

Semantic role labelling(SRL)(He and Liu; 2018) is concept that is quite familiar in the field of natural language processing. It is a way to represent the words in a format for identifying the roles, action, place, time present in the content. SRL is a schematic representation of text data which aids in structuring the content which will eventually aid in machine processing. A machine cannot distinguish between a name, place or animal present in a text content. SRL techinque helps by adding a structure into the text content. Past researches of story comprehension(Chaturvedi et al.; 2017) has shown methods for studying the pattern of sentences in the content in which models are trained to learn the pattern of events in sentences to predict the next sentence. These researches also added extra features to improve the accuracy of the model.

The event feature was modelled based on the SemLM model as proposed by Peng and Roth (2016). In this model, the events will be represented by frames(Baker et al.; 1998) extracted from the content using the FrameNet lexicons. The model was represented as combination of frames and discourse markers. Frames(Chaturvedi et al.; 2017) signifies the context sense of the sentences and thereby giving the event. The following sequence is an illustration of the representation of the SemLM approach:

$$[f_1, dis_1, f_2, o, f_3, o, dis_2, ..., o] \tag{1}$$

This representation is concatenation of frames, discourse markers and period. The research explored a variety of techniques to model the data. Skip gram model(Peng and Roth; 2016), continuous bag of words, ngram and log-linear model. The log linear model was the most effective model out all. It included three vectors in its features: Target, bias and context vectors. This probabilistic model(Peng and Roth; 2016) of frames, period and discourse markers unlike the bag of words model keeps track of the context in the sentences

2.3.2 Frames for context

A unique way to identify the context in text data would be to identify the frame of a sentence. Frames(Baker et al.; 1998) are word representation of a sentence. In a sentence, it is associated with a specific frame evoking word in the sentence. that implies the context of the sentence. For instance for the sentence "Tom is riding a bike" the frame would be 'Operating_bike' and the frame lexical unit would be 'ride'. This way of representing sentences in the form of one word helps in extracting the context of the sentences and also reduces the size of the data thereby removing unnecessary noise words.

The Frame evoking words are refereed to as the Lexical Unit. A frame would have multiple lexical units, depending on the context of the sentence, the lexical unit would vary but the frame would still be the same implying a fixed event. It is this concept that is central to the idea of context identification using frames. As mentioned in the earlier example, ride.v is the lexical unit(LU) that triggered the frame 'Operating_Vehicle'. But ride does not necessarily always trigger Operating_Vehicle alone. The LU could belong to multiple frames.

2.3.3 FrameNet

FrameNet(Chaturvedi et al.; 2017) is a lexical database which is a collection of objects known as the frames which has associated lexical units and frame elements. This interlinked data can used to build models that can understand the context of discussion in text. FrameNet consists of mainly three components: (Baker et al.; 1998)

- Frame Database
- Lexicon
- Annotated Sentences

FrameNet is a structural format of language that has a huge collection of data associated with the schematic representation of language. FrameNet is a collection of frames and the words associated with it known as the lexical units. It also holds the frame element and frame relations data. For example, 'Operating_Vehicle' frame uses 'Motion' frame and the frame elements belonging to 'Operaing_Vehicle' frame are driver, vehicle, path, distance, route, duration, operator etc. These frame elements can grouped together to form a frame elements group(FEG)(Baker et al.; 1998) for schematically representing a sentence. A frame in general provides all information that could surround the context in a sentence.

2.4 Transformation of text into high dimensional space data

Text data cannot be processed as data in its raw form cannot be used for computations. A vector value has a magnitude and direction. A numeric value can be represented on a dimensional space but a text data cannot be represented in any format that can used for computation. So it is important to understand the necessity of the methods to transform the text data in a way that would hold the idea otherwise known as the features that would decide the classification or clustering factors.

Natural language processing(Llorens; 2018) techniques can used to transform the data into features which can eventually be used for executing the tasks aimed at understanding the language. In an unsupervised algorithm the documents are modelled based on similarity in the input feature data. In a supervised algorithm, the output data is used to analyze the pattern in the input data.

One of the most popular method is the count vector representation which forms a matrix of the count of words in every document. The matrix dimensions will be the vocabulary and the documents. This model will give a numeric representation of the documents and hence aids in finding similarities between documents by the count of specific words within a document.

2.4.1 Clustering data

The data that are represented in the numeric format can be clustered or classified using the known algorithms such as KMeans(Krishna et al.; 2012) or KNN. k-Nearest Neighbour(kNN) (Cai et al.; n.d.) is generally used for supervised clustering while KMeans is used for unsupervised clustering. As mentioned earlier, kNN being a supervised algorithm analyzes the data based on the output labels, the classification is done by finding the distance between the data point and its nearest cluster. While KMeans algorithm analyzes the similarities between the data points and clusters into different clusters by analyzing the closeness between the input data.

For the purpose of this research KMeans clustering algorithm will be considered for clustering as the data does not have any labels and the aim of the research is too explore the existence of specific classes present in the text data.

2.5 Why cluster the story plots?

Story plot classes specifies the theme and outline of the story. Understanding a category helps understand the elements that could be present in the story. The book 'Seven basic plots'(Booker; 2004) mentioned that there are seven categories of movies in which all stories should fall under.

- Rags to riches - In rags to riches the protagonist would be at a point where he or she is at successful point in life and then ends up losing it which will be the conflicts introduced in the story. The story would describe the events that he or she goes throuhg in order to restore balance.

- Quest - The story would describe a journey the protagonist takes in order to search for a place, thing or even a person. The story would involve the conflicts the protagonist comes across but at the end finally accomplishes the task.

- Voyage and Return - In this class, the story would involve a protagonist who would undergo a series of conflicts or probably go on a quest and then finally return with new experience.

- Comedy - Comedy class does not have any specific structure as it the other classes could overlap with the class as well. Most romantic movies also would fall under this category.

- Tragedy - In this class the protagonist would face a downfall by the end of the story.

- Overcoming the monster - The story would have a protagonist trying to overcome a monster or a villain.

- Rebirth - A protagonist goes through a series of conflicts and finally resolves the conflicts along with undergoing a change in his or her character a bit, making the protagonist either a better or worse person in life.

Each of these classes have specific story structure. The fact that some of portion of the story structures could overlap, for example, the quest plot and the voyage and return could overlap as in both cases the protagonist would face conflicts on a journey. But both the stories are different, it is facts like this that distinguishes a story from another and it is such overlapping content that may miss out the uniqueness of two stories.
The importance of this kind of a model to classify raw story plot into classes that has specific story outlines is that it adds a label to the story that would be reduce the complexity of understanding the content in the data. This is the same principle followed in FrameNet where the frames are connected to lexical units and frame elements and frame element groups. Connecting all relating information structures the data which can be processed by machine. For instance a rebirth class signifies that the story content would consist of elements like 'Protagonist Change', 'Cause of Conflicts','Character before change', 'Character after Change'.

This research explores a method to structure story content in order to be able to process the content in a machine which can eventually be used for business applications of movie recommendation systems. Such models can also be beneficial for sentence comprehensions where the model tries to predict the next event in a story. The idea of a story is to invoke curiosity in the minds of the viewers. A model that can identify the pattern of events in a structure can very well be used to identify the type of events a viewer is fond of. This is not only beneficial for a recommender models but also adds value to the recommendations made that is more satisfactory to the end user.

3 Methodology

The research implementation is following a KDD (knowledge discovery in database) process(Fayyad et al.; 2013). The implementation of this research focuses on analysing large amount of data to identify the existence of an unknown pattern in the data. KDD methodology follows a framework that concentrates on the steps involved rather than the team management. It involves:

- Selection - The plot summaries for the study is taken from the CMU Movie summary corpus(Bamman et al.; 1996). The story summaries forms the primary data for analysis in the study.

- Pre-processing - In the pre-processing stage, the data is cleaned of all null values. The cleaned data then undergoes feature extraction using a frame classifier model.

The frame classifier model is built using a collection of frame annotations which is further processed for balancing the data to consider data with only 50 rows for every frame label.

- Transformation - The pre-processed plot summary data is then transformed into frames using the frame classifier model. The frame sequence data is then further transformed into vector values using a count vectorizer forming a matrix representation of the word and its corresponding counts in each document.
- Data Mining - The vector representation is then clustered using k means clustering which is configured to cluster the data into 7 clusters.
- Evaluation - The clusters generated are then evaluated for authenticity of the clusters extracted. The evaluation is done using Elbow test and Silhouette test.

3.1 Pre-processing

The plot summaries taken for analysis was first analyzed for null values to remove the null cases which would impact the further transformation processes. The raw data was transformed into frames in the transformation phase. To prepare the data for this transformation a frame classifier model was built using annotated sentences with frame labels. The data is extracted from the FrameNet lexical database. Number of frame annotations for every frames are varying throughout the lexical database, so the data was balanced by considering only 50 annotations from every frame. Using this data, a Linear SVM(Apostolidis-Afentoulis; 2015) model is trained to classify a sentence into its associated frame. The trained model showed a 67% accuracy. The frame data was filtered for only frames associated with verb meanings because plot summaries are talking about actions or events in a story so the associated frames are triggered by verbs in the sentence. For example, "Tom rode a bike" would be belonging to 'Operating_Vehicle' frame which is triggered by the word 'rode' which in its root form would be 'ride'.

3.1.1 Support Vector Machine

Support vector machine(SVM)(Apostolidis-Afentoulis; 2015) is an old and popular machine learning algorithm. SVM with linear kernel separates the positive labelled data from rest of the data which is considered as the negative labelled data with the largest separation margin hyperplane. The idea of largest margin for the hyperplane is considered for the classification due to its advantage of being more resistant to noise data. One of the main features of SVM is that it has high accuracy.

For this research, a linear kernel was chosen in order to classify the annotations labelled with assoiciated frames. SVM algorithm was chosen to classify the annotations as the the total number of classes were 374. The annotations were associated with the frames based on a specific lexical unit(Baker et al.; 1998) and the surrounding context. This required a highly accurate classifier as compared to other algorithms.

3.1.2 Numeric representation of Frames

The extracted frames is then converted to numeric form for further clustering. The numeric representation is obtained using a count vector(Krishna et al.; 2012) which forms a matrix representation of every documents and the count of every word in every document. The dimension size will be equivalent to the entire vocabulary size of the frames extracted. The count vector method is preferred over the word vector(Le and Mikolov; 2014) to transform the data because word meaning would not be very effective in such cases of transformed frame data. As a frame and its surrounding frames will not be dependent on each other like a word and its surrounding words.

3.2 K-Means clustering of frame data

The raw text data transformed into frame sequence is converted to its corresponding numeric representations. The count of every word in a document is represented in the form of a sparse matrix. This sparse matrix is fit into a K-Means(Krishna et al.; 2012) clustering model to cluster the data points into 7 clusters. The algorithm will cluster the data based on the count of specific words in each documents thereby clustering the documents based on similar characteristics. The number of cluster was predefined in the algorithm. Depending on the k value chosen, the clustering improves. At the optimal value of chosen k the errors will be minimized and data will be clustered far from each other.

4 Design Specification

4.1 Dataset

Data adopted for undergoing the research is frame annotations from FrameNet(Baker et al.; 1998) and plot summary corpus data from a published movie corpus data(Bamman et al.; 1996) which is a collection of movie data gathered for a research that was funded by the US National Science Foundation. There are 42303 stories gathered in this dataset. The frame annotations were extracted from and used for training a model to classify text into frames. There are 201424 frame annotations available in the lexical database. The movie data was transformed into sequence of frames using the frame classifier.

4.2 Requirements

The entire research was implemented using Python libraries and functions. The growing community of Python enables data analysis to be performed at a more accurate and elaborate manner. The minimum requirements for executing this research is:

1. Python version 2.7
2. FrameNet Data version 1.7
3. 8GB RAM CPU

4.3 Process Flow

The process design follows the KDD process. The data selected is preprocessed and transformed and then fit into a model. The model is evaluated for optimal k value. The optimal k value is then used to cluster the data into the appropriate clusters.

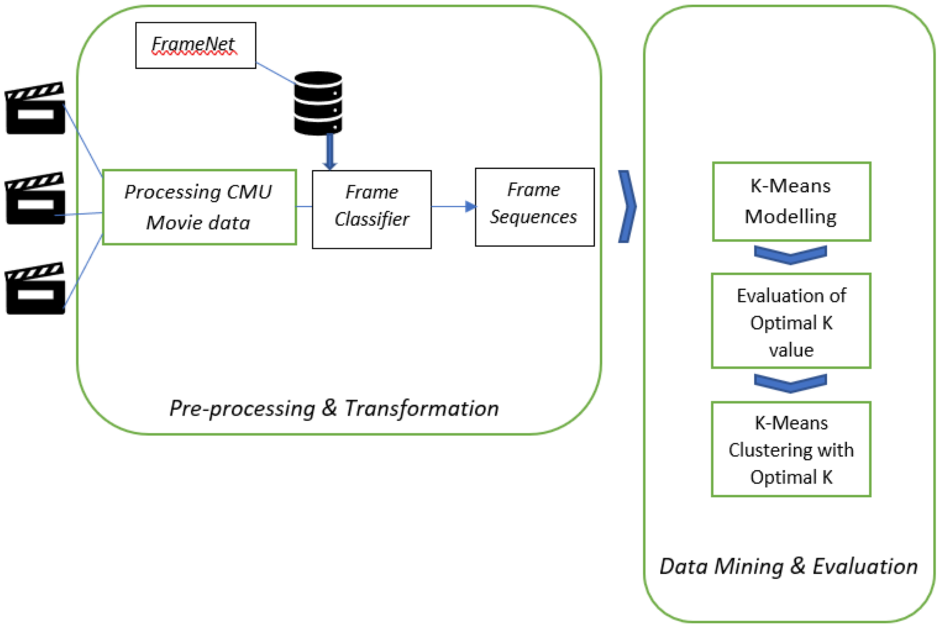
FrameNet
Processing CMU Movie data
Frame Classifier
Frame Sequences
Pre-processing & Transformation
K-Means Modelling
Evaluation of Optimal K value
K-Means Clustering with Optimal K
Data Mining & Evaluation

Figure 3: Process flow

5 Implementation

The research was implemented integrating elements that are specific to Natural language processing and statistical modelling. The implementation involved integrating FrameNet data with the Movie dataset considered for the research. The results obtained from the implementation was then further evaluated to verify the results. The implementation will be divided into three stages

1. Frame Classifier
2. Frame Identification in Movie data
3. Clustering

5.0.1 Frame Classifier

The frame classifier is modelled using the annotations data present in the FrameNet(Baker et al.; 1998) lexical database. The annotations along with its frame labels are extracted from FrameNet and used for training the classifier model. A linear SVM(Apostolidis-Afentoulis; 2015) classifier from the SciKit library of Python was used to build the model.

5.0.2 Frame Identification

The CMU movie corpus data taken for the analysis is tokenized and classified into frames using the frame classifier model(Linear SVM). The model classifies the tokenized sentences into a sequence of frames. This sequence of frames is the transformed data that is now ready to be clustered.

5.0.3 Clustering

The sequence of frames obtained for every story plot is further transformed into a sparse matrix using a count vectorizer so that the data can be represented in high dimensional space. The sparse matrix is trained using a K-Means(A Syakur et al.; 2018) model with a predefined k value. The K-Means model function was obtained from SciKit library in Python.

6 Evaluation

The idea of the research is to prove the existence of structure in story scripts in the form of 7 clusters. The clustered result obtained from the implementation needs to be verified for the validity of the clustering. The clustering is done based on predefined k value. Based on the correct value of clusters chosen, the clustering is done more accurately and the data points are clustered far from the nearest cluster.
The evaluation steps are as following:

1. Evaluating best fit frame Classifier Model
2. Elbow test
3. Silhouette Analysis

6.1 Frame classifier evaluation

6.1.1 Naive Bayes with no POS tagged frames

The frame annotations with labelled frames data obtained from FrameNet was used for modelling the Naive Bayes classifier. The number of labels considered for the model was 1014 frame labels. The model was trained with the data to obtain a model with an accuracy of 25% accuracy which does not seem like a model good enough to classify the labels.

6.1.2 Naive Bayes with POS tagged frames

The frame annotations were POS tagged and filtered for only verb tags. All frames that were only verbs were identified by POS tagging and out of the filtered data, only the words that were tagged as verbs were considered in the data. The data was used for modelling the Naive Bayes classifier which showed an accuracy of 32% accuracy.

6.1.3 Linear SVM with no POS tagged frames

Frame annotations with labelled data was then used to model a linear SVM(Apostolidis-Afentoulis; 2015) model to classify the frames data with 1014 frame labels. The model showed an accuracy of 61% which was an improvement to the Naive Bayes classifier.

6.1.4 Linear SVM with POS tagged frames

The frame annotations extracted from the FrameNet was filtered for only verb frames by POS tagging and normalizing the extracted annotations data from FrameNet, so as to classify the sentences to frames associated with verbs. The frame labels The model showed a further increase in the accuracy to 67%.

Frame Classifiers		
Methods	Naive Bayes(%)	Linear SVM(%)
Without POS Tag	25	61
With POS Tag	32	67

6.2 Elbow test

Elbow test(A Syakur et al.; 2018) is an evaluation technique used for finding the optimal number of clusters required to cluster a dataset. The method is implemented by plotting a graph of sum of squared errors(SSE) vs number of clusters. The SSE(A Syakur et al.; 2018) is calculated for every sample value of k within a particular range and the the curve is observed to drop with the increase in number of clusters. The curve will substantially drop until a specific value of k after which the curve will stabilize and the sum of squared errors will not vary by a huge margin. This k value is the optimal number of clusters.

6.2.1 Elbow test with imbalanced data

The frame classifier model classified the sentences into sequence of frames. The obtained frames was then clustered into data points using k means algorithm. The data was clustered for a range of k values to analyze the variation in error. The variation was plotted against the number of clusters within the range 1-13 to obtain an optimal k value of 3 clusters. The total data considered for clustering is 42303 stories.

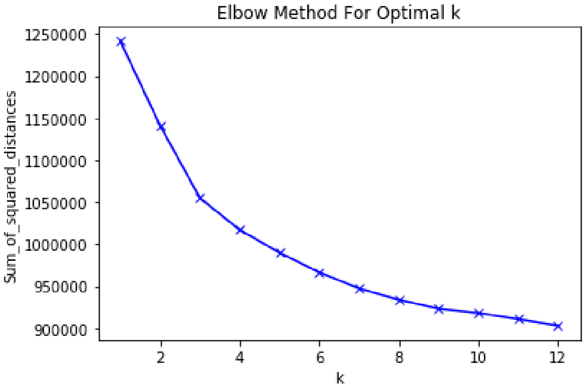

Figure 4: SSE vs Number of clusters

In Figure 4, it can be observed that at k=3 the curve has made an angle and the significant change in square distances has dropped. It is after k value of 3, the elbow curve has significantly dropped.

6.2.2 Elbow test with balanced data

The number of sentences in certain story plots were too small causing an imbalance in the data for modelling leading to clustering data based on size of data in documents. The data was filtered with criteria of a minimum value of 20 sentences in a story. The extracted data was vectorized and clustered using K means algorithm. The error variation was plotted against the number of clusters within the range 1-13 to obtain an optimal K

value of 6 clusters
In Figure 5, it can be observed that at k value of 6 the curve has flattened and stabilized.

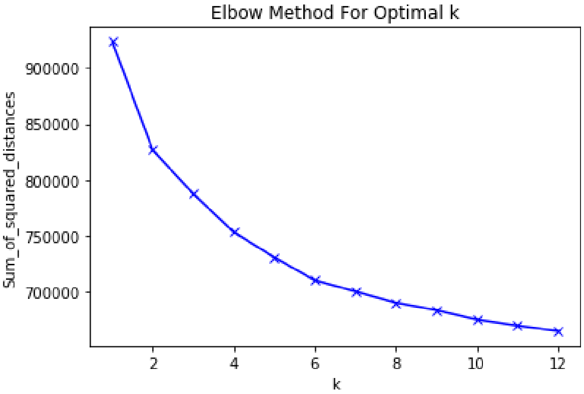

Figure 5: SSE vs Number of clusters

For all k values greater than 6, the variation in squared distances are not substantial variations.

6.3 Silhouette Analysis

Silhouette analysis is a popular method used for analyzing the best fit model for clustering algorithms. For unsupervised algorithm, since the labels are unknown, a value has to be assumed when clustering. On varying the k value the variation in error can observed. The silhouette analysis calculates an index value which determines the clustering strength. In silhouette analysis(Thinsungnoena et al.; 2015), a silhouette index is calculated using the generated model. The value of silhouette index ranges between -1 and 1. These values determines the validity of the clustering. Table 1 indicates the significance of the index values.

Table 1: Silhouette Index Values.

Silhouette Index	**Interpretation**
1	Data points are clustered far away from nearest clusters
0	Data points are close to the nearest cluster
-1	Data points are clustered in the wrong cluster

Index value of -1 indicates that the data points are clustered in the wrong cluster. Index value of 0 indicates that the data points are close to the nearest cluster. Index value of 1 indicates that the data points are clustered far away from nearest clusters. These index values describes the clustering strength of the model and also indicates the variation in distances between the clusters based on the number of clusters chosen.

The analysis results shown in Table 2 shows that index values all are positive indicating that the clusters are all apart from each other and the data points have not been wrongly clustered. On closer observation it can be noticed that only beyond k value of 6 the index value variation tends to drop. Although k values between 2,5 does have a good index values indicating good clustering strength, there is too much variation between this

Table 2: Silhouette Index Values

K	Silhouette Index
2	0.709
3	0.235
4	0.162
5	0.157
6	0.055
7	0.049
8	0.044
9	0.037
10	0.034
11	0.025
12	0.018

range. These results indicates that the optimal number of clusters in the movie data is 6. The data has 6 types of data having a 6 types of frame combinations.

6.4 Discussion

The results from the evaluation of the clusters have brought light to the fact that there is a structural pattern in the story plots. Table 3 shows the cluster labels as per data observed in the clustering results. It was observed to be classified within 6 clusters out of which 5 clusters were clustered according to the story content and 1 was clustered based on the cluttered data. Keeping the cluttered data cluster aside the remaining 5 clusters were clustered based on the content of the story into Voyage and return, quest, overcoming the monster, rebirth and rags to riches.

Majority of the stories were clustered under cluster 1,3 and 5. All stories which involved a character going around a series of events and then finally returning to a scenario as similar to that of the beginning but with an additional positive experience or theme fell under cluster 0 labelled as Voyage and return plot. All stories with a character undergoing a transformation in character after a series of events leading to mostly a positive character fell under cluster 4 labelled as 'Rebirth'.

Table 3: Clustering result

Cluster	Class
Cluster 0	Voyage & Return
Cluster 1	Quest
Cluster 2	Random Data
Cluster 3	Overcoming the monster
Cluster 4	Rebirth
Cluster 5	Rags to riches

Cluster 2 consists of cluttered data i.e. data that is not complete and does not make any sense. These stories are a mixture of outlines but falls under similar characteristics of cluttered data. Hence the results shows that the data was clustered into 5 clusters appropriately.

7 Conclusion and Future Work

The study has identified the factor that story plots do have a structure in text content. This research has focused mainly on the event pattern in the sentences to cluster the data which is an important factor to analyze a pattern in text data. The past researches based on topic word classification focuses only on the variation in the noise(topic words) between different classes. The events identified from the sentences analyzes the pattern of the structure in the story plots which is the key findings in this research. The stories were appropriately clustered into the 5 labels out 7 labels mentioned in 'the seven basic plots'. The outline classification model sets a base for further analysis of the text data.

The research is still open for observing for factors that is responsible for clustering the data into 7 plots. The outline classification needs to be further analyzed to extract elements(conflict cause, journey path, character change etc.) from the plots. These elements aids in improving the machine learning model to understand the data better. The research was implemented basing the idea of FrameNet and the three act structure(Morris; 2008) which explains the three phases in a story and the observation made in this research was that exposition, rising action and the climax phases was impacting the clustered results. Further exploration needs to be done on how the frames(Baker et al.; 1998) can be utilized for segmenting the story content into a three act structure model.

8 Acknowledgement

I would like to express my special gratitude to Dr.Anu Sahni for her supervision, guidance and support through out my research. I would also like to thank Dr.Vladimir Milosavljevic for his valuable inputs that were beneficial for my research study.

References

A Syakur, M., K Khotimah, B., M S Rochman, E. and Dwi Satoto, B. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster, *IOP Conference Series: Materials Science and Engineering* **336**: 012017.

Abernathy, T. and Rouse, R. (2014). Death to the three act structure! toward a unique structure for game narratives.
URL: *https://www.gdcvault.com/play/1020050/Death-to-the-Three-Act*

Apostolidis-Afentoulis, V. (2015). Svm classification with linear and rbf kernels.

Baker, C. F., Fillmore, C. J. and Lowe, J. B. (1998). The berkeley framenet project, *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, Association for Computational Linguistics, pp. 86–90.

Bamman, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). Learning latent personas of film characters, *AI magazine* **17**(3): 37.

Booker, C. (2004). *The seven basic plots: Why we tell stories*, A&C Black.

Cai, Y.-l., Ji, D. and Cai, D. (n.d.). A knn research paper classification method based on shared nearest neighbor.

Chaturvedi, S., Peng, H. and Roth, D. (2017). Story comprehension for predicting what happens next, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1603–1614.

Ertugrul, A. M. and Karagoz, P. (2018). Movie genre classification from plot summaries using bidirectional lstm, *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pp. 248–251.

Fayyad, D., O'Connor, B. and Smith, N. A. (2013). From data mining to knowledge discovery in databases, *ACL* .

Gallo, C. (2016). *The storyteller's secret. 1st ed. London: St. Martin's Press*, St. Martin's Press.

He, S., L. Z. Z. H. B. H. and Liu, G. (2018). Syntax for semantic role labeling, to be, or not to be.
URL: *https://aclweb.org/anthology/P18-1192*

Hoang, Q. (2018). Predicting movie genres based on plot summaries.

Hoeken, H. and van Vliet, M. (2000). Suspense, curiosity, and surprise: How discourse structure influences the affective and cognitive processing of a story, *Poetics* **27**: 277–286.

Kar, S., Maharjan, S. and Solorio, T. (2018). Folksonomication: Predicting tags for movies from plot synopses using emotion flow encoded neural network, *arXiv preprint arXiv:1808.04943* .

Krishna, B. V., Satheesh, P. and Suneel Kumar, R. (2012). Comparative study of k-means and bisecting k-means techniques in wordnet based document clustering, *International Journal of Engineering and Advanced Technology* **1**(6): 1–4.

Lahlou, F., Benbrahimand, H., Mountassir, A. and Kassou, I. (2013). Context extraction from reviews for context aware recommendation using text classification techniques, pp. 1–4.

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

Liu, Q., Huang, H., Gao, Y., Wei, X., Tian, Y. and Liu, L. (2018). Task-oriented word embedding for text classification, *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2023–2032.

Llorens, M. (2018). Text analytics techniques in the digital world: Word embeddings and bias.
URL: *https://arrow.dit.ie/cgi/viewcontent.cgi?article=1157context=icr*

Makita, E. and Lenskiy, A. (2016). A multinomial probabilistic model for movie genre predictions, *arXiv preprint arXiv:1603.07849* .

Morris, A. K. (2008). *It's all a plot: an examination of the usefulness of the popularly accepted structural paradigm in the practice of writing of a feature film script*, PhD thesis, Queensland University of Technology.

Motro, S. (2015). The three-act argument: How to write a law article that reads like a good story.

Peng, H. and Roth, D. (2016). Two discourse driven language models for semantics, *arXiv preprint arXiv:1606.05679* .

Thinsungnoena, T., Kaoungkub, N., Durongdumronchaib, P., Kerdprasopb, K. and Kerdprasopb, N. (2015). The clustering validity with silhouette and sum of squared errors, *learning* **3**: 7.

Yu, L.-C., Wang, J., Lai, K. and Zhang, X. (2017). Refining word embeddings using intensity scores for sentiment analysis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* pp. 1–1.