# Project Proposal

**Group Members(Group No: 20):**

Calida Pereira - 229945
Chandan Radhakrishna - 229746
Rashmi Koparde - 230322
Priyanka Bhargava - 229675
Thi Linh Nham Dao - 224467

**Objective:**

The main objective of our project is to train a machine learning model such that it automatically annotates the Genre of a novel.

**Features:**

We've short listed a number of features which we could use to train our model for the classification task.As of now it's not fixed with regards to which all features amongst the listed, would be useful but as we proceed with the implementation we intend to compare our model with different combination of features and select the ones that outperform the rest.

| **Feature Type** | **Feature Name** |
|---|---|
| Author Embeddings | Author Name. |
| Writing Style | Female Pronoun, Male Pronoun, Personal Pronoun, Possessive Pronoun,Preposition, Colon, Semi-colon, Hyphen, Interjection. |
| Sentence Complexity | Co-ordinating Conjunction, Comma, Period, Punctuation and sub-ordinating Conjunction, Sentence Length. |
| Sentiment | Negative, Positive, Neutral |
| Ease of readability | Flesch Reading Score |
| Plot complexity | Number of characters |
| Lexical richness | Type Token Ratio |
| Title | Book Name |
| Summary | Word2vec or/and doc2vec on entire data |

**Feature Selection and Extraction Techniques:**

1) Filter methods - Information Gain and Chi square
2) word2Vec
3) Node2Vec
4) Doc2Vec
5) IDF and Bag of Words.

**Machine Learning Models:**

Classifiers that'll be considered while modelling: Support Vector Machine, Multinomial Logistic Regression and Multinomial Naive Bayes.

**Evaluation Techniques:**

1) Precision
2) Recall
3) Micro averaged F1 score

**References:**


1) https://github.com/suhitaghosh10/fictionRetrieval/blob/master/ICHMS2020_paper_42.pdf
2) http://cs229.stanford.edu/proj2015/127_report.pdf