SENTIMENT ANALYSIS FOR MOVIE RECOMMENDATIONS

HILARIO JUNIOR NENGARE(174682)

Professor Dr. Zafer Erenel | EUROPEAN UNIVERSITY OF LEFKE | TRNC

INTRODUCTION

- In this project the main aim is to achieve a movie-suggestion system which is optimal and personalized.
- With Sentiment Analysis we're able to mine for user opinions, felts and sentiments towards a certain movie and then use that input to create a very competent movie recommendation system.
- Sentiments in this regard are like a thumbs up or thumbs down in older movie rating systems.
- Different users can share different sentiments towards a movie, and it is these reviews that we look at and then classify in accordance with positive or negative polarity.
- In order, to provide movie suggestions that align with user sentiments.

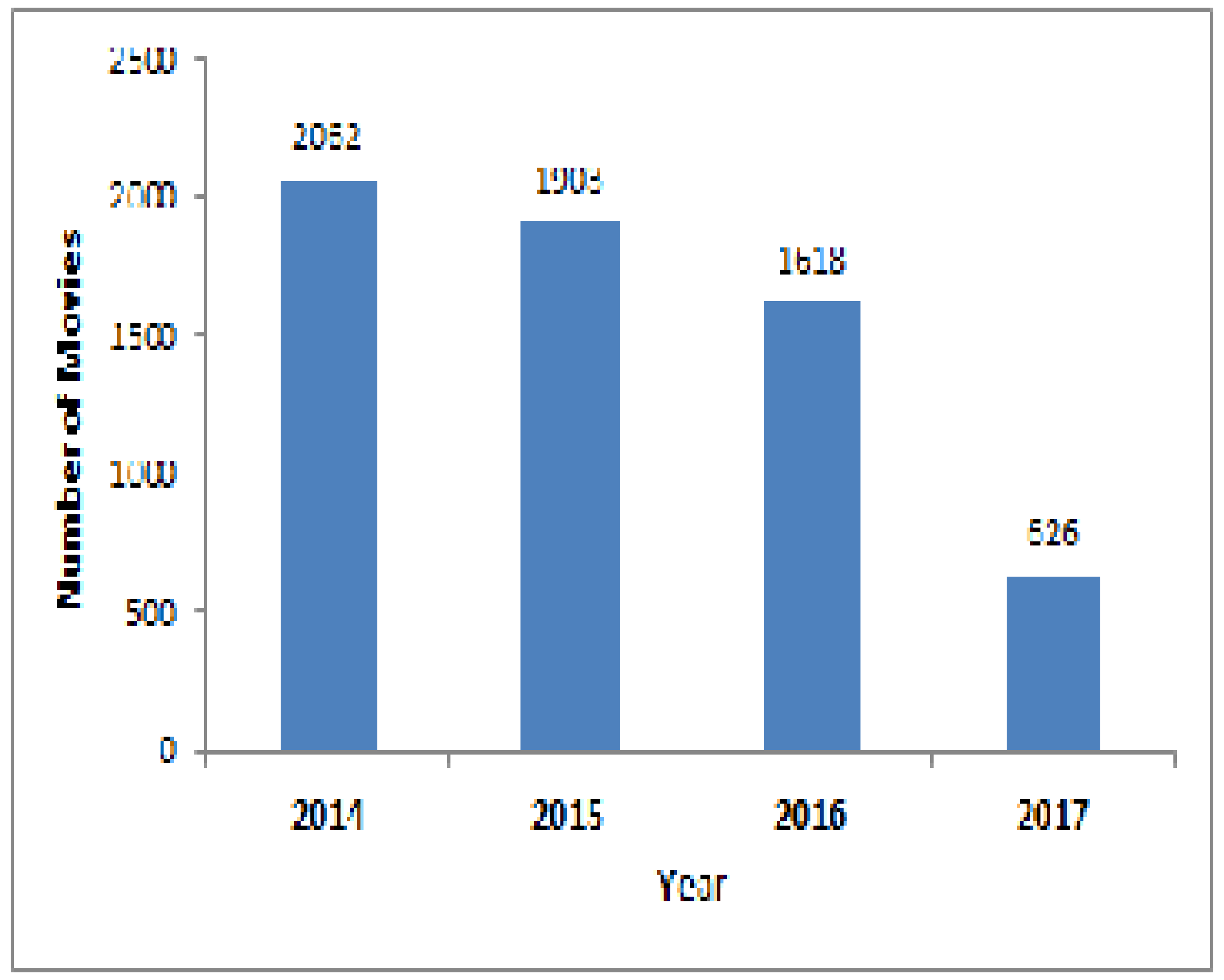
GOALS

- **Viewer engagement** – with personalized movie recommendations, we achieve viewer traffic and manage to retain viewers.
- **Context-based movie recommendations** – each movie pitched to the viewer should be domain-compliant and contextualized to user sentiments.
- **Improved content Discovery –** this is what gears this project as we want to bring movies the user would love but would not have found out otherwise.
- **Adaptive movie recommender** system for users which enhances user retention, by automatically knowing what the user prefers and in what case they prefer it.
- **Increased satisfaction** when users receive recommendations that align with their preferences and sentiments, they are more likely to enjoy the recommended movies.

MODULES

- **Data collection** – A dataset with movie reviews from 2014 – 2017 is to be used due to the rarity of reviews that are social media-based on older movies (< year 2014).
- 45 % of the movie reviews is from movies within this period from 2014 – 2017.
- Which is 20% of TMDB database alone.
- Distribution of these movies is denoted in the graph below

Fig 1: Distribution of movies per year:



- **Data cleaning** - An obtained review dataset is decoded into aspects during training, using the techniques unigram – if it's one word, bigram-if it's two words and trigram if it's three words.
- In order, to circumvent obscure data within our movie dataset say reviews published from different demographic areas, we use both unigram and bigram on Pragmatic features to determine the context of certain text and punctuation.
- In data cleaning module we extract useless features and symbols, such as is denoted on fig 2.

Fig 2: Example of noisy and uninformative data :

NOISE	EXAMPLE
Stop Words	a, and, the, after, am
Lemma	Serve, served, serving
URLS	www.restaurant.com
Filtration of Repeating Aspects	Haapppyy, helloooo
Special Characters	!,#,$,%,_

MODULES (cont.)

- **Feature extraction -** This module involves the extraction of encapsulated text features that are most essential from our movie review data set.
- Removal of punctuation in order to extract essential features is done in this module.
- Many processes cater for this module such as
- **Bag of words**, a process I will use for feature extraction on the dataset.
- For each sentence, a vector is created
- e.g vec = {“this”, “movie”,”is”,”good”} and the size of the vector is 4 hence is represented such as _vec = [1, 1, 1, 1].

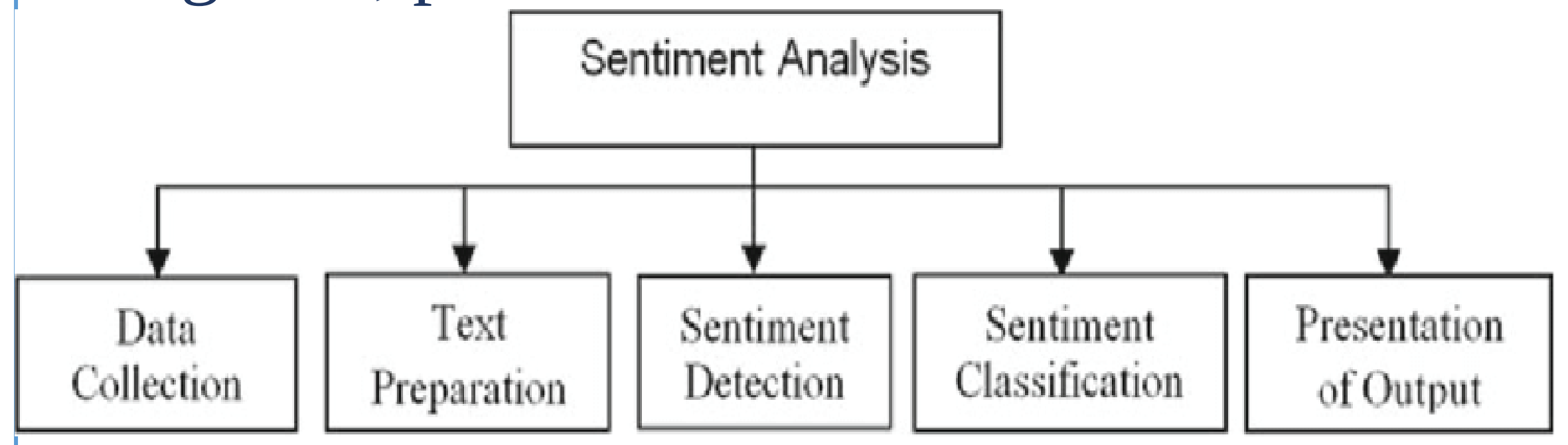
Fig 3: Illustration of Bag of words

Word	this	Movie	is	good
Frequency	1	1	1	1

- **Subjectivity Classification** - Determination of subjective to objective data in the movie review dataset. Distinguishing between subjective and objective reviews.
- Subjective reviews have personalized points of view and objective are based facts and they do not contain bias.
- **Spam Detection -** content-based features where repeated phrases, particular words and irrelevant information is identified as a pattern indicative of fraud/spam.

MODULES(cont.)

- **Polarity Detection -** In this module classification of reviews into negative, positive or neutral is done.



- A machine learning or deep learning model for polarity detection is selected. These include Naive Bayes, SVM, logistic regression and neural networks.
- The selected model is then trained on the labelled data that was attained. I use the extracted features and other corresponding labelled data to train the model.

CONCLUSION

- Sentiment analysis is a very relevant field of NLP and hence there loads of previous findings and research pertaining to this field.
- Its popularity ensures that there are a lot of resources such as libraries and datasets to build upon.
- It can be applied to other fields other than movie sites that use recommender systems and dependence is upon user feedback.
- Below **ethics** ought to be **enforced**:
- It is imperative to protect user data, ensure secure means of interaction, openness and inclusive usage for all people.

REFERENCES

[1]: Sentiment classification using machine learning techniques (2018) Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan retrieved from https://www.cs.cornell.edu/home/llee/papers/sentiment.home.html

[2]: Extracting Lexically Divergent Paraphrases from Twitter by Wei Xu. Retrieved from https://dblp.org/pid/32/1213-4.html.

[3]: Feature selection in machine learning: A new perspective by Cai Jie, Luo Jiawei, Wang Shulin, Yang Sheng. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0925231218302911?fr=RR-2&ref=pdf_download&rr=83b0e8b219245a99

[4] Features extraction retrieved from https://deepai.org/machine-learning-glossary-and-terms/feature-extraction#:~:text=What%20is%20Feature%20Extraction%3F,losing%20important%20or%20relevant%20information.

[5] Sentiment analysis and subjectivity by Bi Liu(2010). Retrieved from https://www.researchgate.net/publication/228667268_Sentiment_analysis_and_subjectivity.