EUROPEAN UNIVERSITY OF LEFKE

FACULTY OF ENGINEERING

Graduation Project I

SENTIMENT ANALYSIS FOR MOVIE RECOMMENDATIONS

Hilario Junior Nengare

174682

Machine Learning

- Machine Learning is a network of models that take input data to predict future inputs.
- It is a branch of Artificial Intelligence which is focused on the implementation of models and mathematical algorithms to make models learn from datasets.
- The goal is to implement models that can learn on their own or adhere to mediated learning.
- To improve models, we expose models to more data sets and improve feedback loops and so forth.
- Although Machine Learning is a subset of Artificial Intelligence, it is comprised of its own subsets; one of which is the focal point of this paper – SENTIMENT ANALYSIS.

Supervisor

Dr. Zafer Erenel

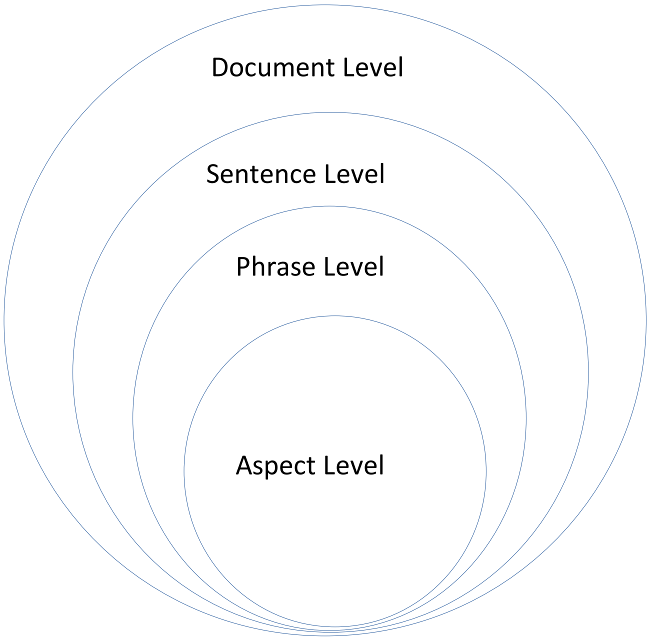
Publish Date

25-12-23

Table Of Contents

1.Introduction

Sentiment Analysis and Its Levels



Document level Sentiment Analysis

- Document level analysis is carried out on a single document
- It's not used as often as other levels of sentiment analysis.
- Document level sentiment analysis is usually applied to chapters and/or pages of books to yield a single polarity.
- With document-level analysis, domain-specificity produces high accuracy.
- The feature vector must be constrained and very specific to a particular domain.

Sentence level Sentiment Analysis

- Within this level; analysis of single sentences is done, and a polarity is appointed.
- Documents carry sentences with different sentiments associated with them.
- A larger training dataset is used, and more processing resources are needed.
- These sentence-based polarity can be used independently or aggregated to yield document-based polarity.

Phrase level Sentiment analysis

- Phrases contain either single aspects or many aspects.

- Analysis is carried out on those aspects.

Finally,

Aspect level sentiment analysis

- In this level we analyse every word and assign a polarity to each word.
- At this level analysis becomes more insightful and granular hence I will employ this level throughout the project.

1.1 Problem definition

- Sentiment Analysis also known as emotion AI or opinion mining, is a way in which we extract, investigate, find traces of affective states and intent behind text.

- Sentiment Analysis is applied in user and/or client written feedback e.g. movie reviews (which will be my focus in this project), social media, survey material and so forth.

- Sentiment Analysis is a process in which we can classify text using reviewer sentiments. This includes positive polarity, negative polarity and neutral.

- For instance, we could have a data set with user reviews on a certain movie

Table 1: An example dataset of user reviews

Viewers	Review	Polarity
User 1	I hate this movie.	negative
User 2	I love this movie.	positive
User 3	I don't love it, but I don't hate it neither.	neutral

- As you can see, different users can share different sentiments towards a movie, and it is these reviews that we look at and then classify them in accordance with positive or negative polarity, sometimes we come across sentiments with weak polarity such as user 3's sentiment, that would be very difficult to classify.

- Sentiment Analysis basically figures out if text is positive, negative or neutral and then we can device a recommendation system from thereon.

Sentiment Analysis and Natural Language Processing

- Natural Language Processing is an interdisciplinary field of Computer Science and Linguistics which is primarily focused on computer to human interaction and vice versa.

- It enables computers to dissect text corpora into well tagged and meaningful input or output.

- NLP is comprised of algorithms and models that are used for generating human-like text or analysing bodies of text and "understanding" bodies of text to produce output or input.

- Now, there are many tasks of NLP that I would be deploying in this project in order to carry out proper Opinion Mining on Movie Reviews.
- These tasks include;

  - ✓ **Text And Speech Processing** – which involves many subtasks like Optical Character Recognition, Speech Segmentation, Speech Recognition, Text to Speech et al. My primary focus will be on Word Segmentation also known as tokenization in which I will be separating blocks of text into single words.
  - ✓ **Morphological Analysis** – Processes like Lemmatization in which one reduces a word to its dictionary form known as a Lemma, stemming which is the same as Lemmatization but does not deflate a word to its dictionary form but to its base form.
  - ✓ **Syntactic Analysis-** This is where we analyse the grammatical set up of a sentence using Grammar Induction process, sentence breaking process and parsing.

- There are other subtasks pertaining to NLP, but the question that all this pose is **what Does NLP have to Do with Sentiment Analysis?**

- Well, the subtasks I listed form the foundation of Sentiment Analysis. I plan to use the subtasks to extract features from text data and then determine the sentiment beneath.

- This will be done with machine learning classifier models that I will train on labelled movie review datasets to perform sentiment analysis.

- Natural Language Processing engulfs all the tasks that are tied to computer science and linguistics whereas Sentiment Analysis is the specific application of Natural Language Processing techniques to figure out the emotion behind text corpora.

1.2 Goals

- What I am gunning for in this project is to exhibit the impact sentiment analysis has or can possibly have in the foreseeable AI future.

- I will be focusing on a movie reviews' dataset, and then weed through verbatim to tag negative, positive and neutral affective states.

- In turn I intend to attain acute knowledge of Natural Language Processing which I will use within my career as a Software Engineer.

- And for users, a personalized experience and movie suggestions aligned with their preferences, mind you this can easily be spread out to other recommender systems also.

- Concisely, I aim to use Sentiment Analysis in the movies domain to attain:

- **Personalised movie recommendations**
  - Improve viewer space by recommending movies that are tailored towards the viewer's sentiments.

- **User Engagement**
  - Offering personalised recommendations tend to enhance viewer engagement as well ensuring user traffic of sorts.

- **Context**

  - Very vital to offer context-aware recommendations to users. Context-aware recommendations will align with the user's mood state and recommend perfect movies.

- **User Preferences**
  - We can track nuanced preferences that includes genres and actors, and we can base these preferences on emotional context.

- **User Feedback**
  - In any product user feedback is such a prolific aspect of product life cycle as we can improve something from user feedback alone.

- **Improved Content Discovery**
  - We can recommend movies users did not explicitly consider, but they will be based on their emotional criteria.

2. Literature Survey

- There is a plethora of research with findings and challenges pertaining to Sentiment Analysis for Movie Recommendations.

- Research suggests that box office success of movies can be predicted by analysing sentiments within viewer reviews.

Other Projects

- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan produced a study to address the difficulties faced in classifying documents using sentiments in contrast to topic-based classification of documents [1].

- They chose to focus on a movie reviews dataset to clearly distinguish positive to negative polarity and neutral.

- In their research they annotated case studies and applications aligning with topical document classification – classifying documents in terms of its subject matter. Some of these case studies are as follows

- **Source Style Document Classification (Biber, 1988)**
  - This aimed to classifying documents in accordance with the document's source or source-style using statistical methods.
  - It was served as a methodology to discern patterns or variations or trends in the way the document's language was being used.
  - SSDC analysed variations of the writing style and utilised computational mathematics' algorithm to manipulate large text corpora.

- **High-brow vs. popular or low-brow (Monstellar and Wallace 1984)**
  Focus was placed on language style variations, distinguishing high-brow from low-brow.

- **Stylistic Variation in Language (Argamon and Engelson et al. 1998)**
  Extended the study of language-style and contributed to the factors that influence the variations in language.

- **Stylistic Analysis (Tomokiyo and Jones 2001)**

Related to language variations; it was an extended exploration focused on statistical methods used in denoting patterns found in written text.

- **Sentiment Analysis and Language Style (Kessler et al. 1997)**
  Contributed insights of classifying text corpora on the grounds of style and/or sentiment.

- **Subjective Genres (Karlgren and Cutting 1994)**
  - Objective was to determine the genre of the text and included subjective genres. This introduced classification by genre as a new field of research.
  - Finn and others in 2002 further explored subjective genres building upon the work of Karlgren and Cutting.
- **Features and Subjective Language (McKeown 1997, Hatzivassiloglou and Wiebe 2000, 2001)**
  - In 1997 McKeown and Hatzivassiloglou's research was grounded on classifying the semantic orientation of words and phrases.
  - In 2000 Hatzivassiloglou and Wiebe focused on detecting features indicating the usage of subjective language which they utilise for categorising documents by subjectivity.

- **Sentiment Based Classification (Turney and Littman 2002)**
  - Turney and Littman explored sentiment-based classification of documents using unsupervised learning.
  - This introduced unsupervised learning algorithms for sentiment analysis techniques.

- **Knowledge-based Sentiment Analysis (Hearst 1992 and Sack 1994)**
  - Hearst dived into the study that involved classification of documents using models that are influenced by cognitive linguistics.
  - Sack expressed the idea to integrate cognitive linguistics into sentiment-based categorization.

- **Discriminant Word Lexicons (Huettner and Subasic 2000, Das and Chen 2001, Tong 2001)**
  - Huettner and Subasic explored the idea of sentiment-based categorization of text corpora using discriminant word lexicons. Expanding on the notion of knowledge-based sentiment analysis.
  - Tong, Das and Chen in the year 2001 built on the work of Huettner and Subasic where they focused on sentiment-based categorization of text corpora.

- **Unsupervised Learning Based on Mutual Information (Turney 2002)**
  - Turney researched unsupervised learning for sentiment analysis.
  - At this point we see the dawn of current studies being carried out leading to sentiment analysis.
  - There are other baseline experiments exploring the inherent difficulty of sentiment-based classification with comparative analysis of different methods, highlighting the challenges of sentiment analysis.
  - Turney in 2002 figured that movie reviews are the most difficult to classify using sentiments.
  - He reported an accuracy that fell around 65.83% on a dataset of size 120

3. Background Information

3.1 Required software

Below are the tools I will be using in this project

1) **Python:**
   - Python is fit for Natural Language Processing as it resembles English language and pseudocode.
   - Its simplicity is ideal, as one gets to focus more on Natural Language Processing rather than focus on the language's syntax.

- Python offers a big and robust support community, many problems have ready-made solutions on platforms like stack overflow, Quora et al.
- Python offers a plethora of NLP libraries as well as machine learning libraries e.g NLTK, Tensorflow, TextBlob, spaCy, Pytorch, Transformers (Hugging Face), sci-kit learn et cetera.

2) NLTK:

- NLTK is a Python library that works with human language data.
- It is powerful and provides simplistic interfaces to over 50 corpora and lexical resources.
- It has a large support online community.
- NLTK has many libraries that one can utilise for classification, parsing, tagging et al.
- NLTK is compatible with Windows, Mac OS X and Linux.
- NLTK is the top tier library to perform sentiment analysis.

3) JUPYTERLAB:

- It is versatile platform on which one can carry machine learning tasks.
- It enables step-by-step development which is beneficial for xploring datasets and visualisations.
- Its interactivity is an advantage, it gives real-time feedback allowing one to debug their code immediately.

4) MATPLOTLIB

- Matplotlib is a python library that is used for creating visualisations of data.
- It is a perfect tool for one to create plots and charts.
- Will use matplotlib to create visualisations of sentiment analysis results.
- Will use matplotlib to map out distributions of positive, negative and neutral sentiments on a chart.

3.2 Other software

(1) PyCharm:

- Great and lite ide for this project.
- Contains much software to complement transpiling, compiling and debugging.
- Contains great code generation, hence dwindling time consumption.
- Has an integrated version control system and inbuilt CLI.

(2) Git:

- Will employ GIT to track changes in the NLP project.
- Git enables one to roll-back to previous versions in case of discrepancies.
- Git is awesome for prototyping as one can create multiple branches without affecting the main codebase.
- It’s useful where one needs to experiment with various NLP models and algorithms.
- Git can be integrated with Bitb++++ucket.

3.3 Hardware

- **Device:**
    - This software is accessible on any device that connects to the internet and has a Browser installed.

4. Modules

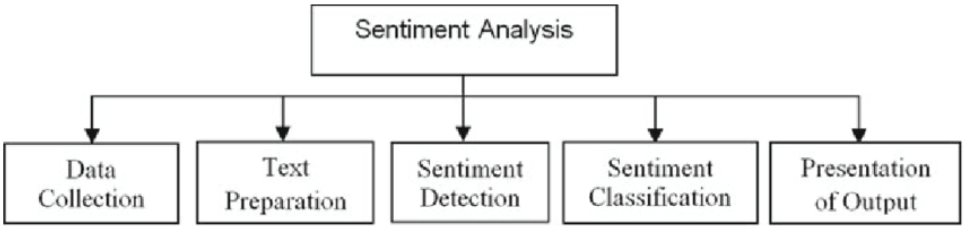


Figure 1: Sentiment Analysis Steps

4.1 Data Collection

- This is one of the most crucial stages in sentiment analysis or machine learning as a whole [2].
- The data that one collects in this stage, is important because this stage almost dictates the success of the project.
- Data can be collected from multiple sources such as social media, customer reviews, videos and surveys.
- There are platforms where we can retrieve ready-made movie datasets from, such as Kaggle, IMDB and others.
- Programmatically one can opt to scrape this data from movie websites such as IMDB but sometimes this can pose to be illegal and against some website's code of conduct.
- In this project I will scrape data from IMDB, The Movie Database, OMDb API and Kaggle.
- In order to carry out the above scraping process, I will use SELENIUM, a python automation-testing software.

4.1.1 Data Cleaning

- In this module there is filtration of inconsistences and correct errors, that might cause problems later.

- This process includes taking care of missing values, correcting inaccuracies, dealing with over the margin data that will not be of any use, and ensuring data integrity.
- The goal of data cleaning module is to improve the quality of the dataset before it is used for either analysis or models.

4.2 Feature Selection

- In order to successfully implement a classification model, one needs to first identify the useful features obtained during data collection[3].
- An obtained review 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.
- Using both bigram and unigram is most effective.

- During feature selection, pragmatic features which refers to words, punctuation and emojis are considered.
- Pragmatic features determine the context of certain text, punctuation directs the orientation and even context while emojis are an expression of human emotion.

4.2.1 Feature Extraction

- Feature Extraction module involves the retrieval of relevant, valuable information from a dataset and directly affects performance[4].
- This module involves the extraction of encapsulated text features that are most essential.
- 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. 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]. Bag of words is evaluated using TF-IDF. Each word occurs once in the document.

Table 2: Example BoW

Word	**this**	**Movie**	**is**	**good**
frequency	1	1	1	1

- **Terms Frequency** is a technique that can be carried out, it is applied to unigrams, bigrams or trigrams.
- Terms frequency has a count to track how many times a word is found in text; the presence of the term is represented as either 0 or 1 with the frequency being an integer.
- **Parts of speech tagging** is another process I will use in this module, also known as grammatical tagging, involves tagging an aspect based on its meaning.
- e.g. “This movie is great” can be tagged as

Table 3: Parts of speech tagging

This	DETERMINER
Movie	NOUN
Is	VERB
Great	ADJECTIVE

- **Negations** is another process within this module, whereby certain aspects such as **not, cannot, neither, never, nowhere, none** can alter the meaning of a phrase and reverse the polarity.

- One might say “This movie is good” which is a positive polarity, but when one says, “This movie is NOT good”, the negation NOT reverts the initial phrase’s polarity to negative.
- At times negations are eliminated as they have neutral sentiment value in a lexicon and no effect to polarity.
- All these processes I will deploy in this module to have better performance in this module.

4.3 Subjectivity Classification

- Determination of subjective to objective data in a piece of text. [5]
- **Subjective** reviews have personalised points of view which are heavily opinionated and based on reviewer emotions.
- The content in subjective reviews is mainly language which is inspired by the reviewer’s perspective, belief and their attitude.
- **Objective** reviews on the other hand are based on facts and do not contain bias.
- In objective reviews there are no personal opinions or emotions but is from observations and verified facts.
- Subjectivity classification aids in classifying polarity – negative, positive, neutral data.
- Models are trained on classified data in this regard – objective or subjective or sometimes separately.

4.4 Spam Detection

- In opinion spam detection, detection of fake reviews is done[6].
- It is NLP module for sentiment analysis where we identify some fraudulent reviews from user opinions.
- In order to maintain integrity of data and credibility it is necessary to perform Opinion spam detection and some of the techniques are
- **content-based features** where repeated phrases, particular words and irrelevant information is identified as a pattern indicative of fraud/spam.
- **Stylometric features** where sentence construction and grammar is used in reviewer text. This may be from automated systems which exhibit consistent writing style.
- Other **techniques are metadata analysis, domain specificity, collaborative filtering and user-behaviour features** which are all effective in spam detection.
- This is helpful in analysing the behaviour of users who post reviews, such as the frequency of posting, the number of reviews from a single user, and the timing of reviews.
- Spam reviews may come from fake accounts or exhibit unusual posting patterns and try to abate that.

4.5 Polarity Detection

- Here are the key steps involved in polarity detection:
- It also known as sentiment polarity detection or sentiment analysis its the most vital module in NLP that determines the affective state of text.
- In this module classification of text into negative, positive or neutral is done.

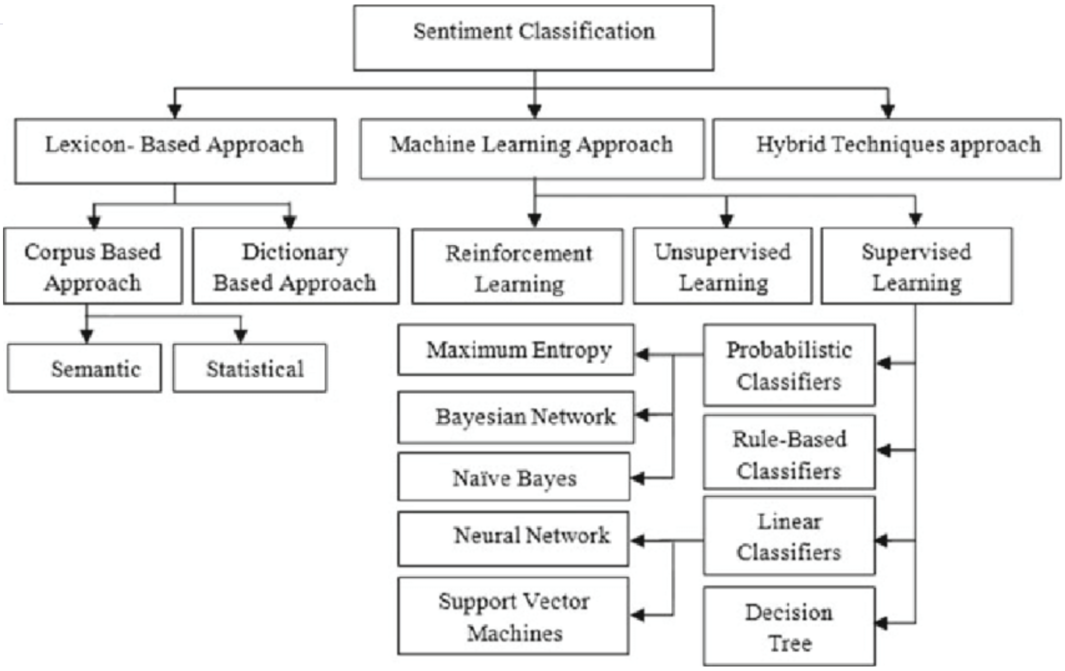


Figure 2: Sentiment Classification

- 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.
- In Polarity detection we execute the sentiment analysis on text corpora and use the attaining in improving movie recommendations.

5. Risk Analysis

a) **Informal Writing**

- Users may use informal abbreviations that may be difficult for models to interpret.
- Users may make use of sarcasm and irony, of which it can be almost impossible to detect and classify.
- Torrents of acronyms are used all over the internet and hence our review dataset might not be void of these.
- *BRB, LOL, IDK, FYI* are some of the acronymic puzzles, that complicate model engineering.
- Sarcasm can be impossible to detect in these cases marginalizing polarity to a faint line.
- Models cannot detect if some sarcastic statements are a pat on the back or a subtle jab.

b) **Language barricade**

- User reviews might be in diverse dialects and hence a threat to accuracy is posed.
- Many languages outside of English have numerous grammatical constructs, strange symbolisms and distinct punctuations enhancing further complexity on models.
- Meaning of certain words/aspects can be acutely different across languages found in the data set.
- Sometimes even words of slang can confuse models posing much complexity, as they can vary across people and languages.
- Keeping up with evolution of dynamic linguistics is a complicated task for models.
- With the language barrier, the models' complexity is at its pinnacle.

c) **Context and Domain**

- Words can make or mar, and this applies to sentiment analysis as well.
- Words are embedded with nuances, that can steer context, and such an example is the word **cool** which may mean okay in social norms and **not warm or cold** in reference to weather.
- Context of text can vary depending on the way its employed and even changing the orientation of that text.

6. Ethics

- **Privacy** – Any data the user shares or uses is kept as such – private. This data is going to be handled in accordance with privacy regulations and will not be used outside of the terms of service.

- **Consent -** Clear statement of how the user's data is to be used in this recommender system and no data is used that is not consented or legal. Clearly communicate how user's data is to be handled and used.

- **Biases and Fairness** – Removal of bias from the dataset, this enhances the integrity of the input leading to fair recommendations.

- **Openness** – Transparency to standards that are used to determine the results of this analysis. Clear communication of the criteria used to influence the recommendations. Users ought to be well informed.

- **Inclusive Usage**– This recommender system ought to be diverse and usable with anyone and anywhere. The system ought to accommodate users from different demographics and cultures.

- **Security** – User data ought to be protected within a robust setting. Protection from spam ware and phishing. Recommender systems often rely on user preferences and user profiles hence its necessary t o safeguard that information.

7. Conclusion

7.1 Benefits

a. Benefits to users:

1. **Context based recommendations -** Users get to watch movies that are specific to their sentimental decisions. Each watch is tailored to the user's sentiments and thus enhancing user retainment and engagement.

2. **Recommendations tailored towards current mood** – Movies can be aligned with the user's current mood. If a user is in a happy state - happy-themed movies are then recommended. Users'

emotional responses will mould their recommendations enabling the system to suggest movies based on the user sentiments.

3. **Personal recommendations –** The system understands the user's emotional response and recommend movies that align with that. The movie catalogue is then tailored towards a personalised taste for the user.

4. **Discovery of new content –** System introduces the user to movies they might have not found otherwise but falls within their preference. Analysing reviews will lead to discovery of other content that aligns with the user's sentiments.

5. **Reduced decision making –** reduced decision making and scrolling to find movies and therefore time saving for the user. Users can focus on a concentrated set of movies/options.

6. **Adaptive recommender** systems for users which enhances user retention, by automatically knowing what the user prefers and in what case they prefer it.

7. **Increased satisfaction** when users receive recommendations that align with their preferences and sentiments, they are more likely to enjoy the recommended movies.

b. Benefits to me:

1. Well, this project is a bit complex and hence, I gain more experience in planning, requirements engineering, data analysis and model training.

2. Artificial Intelligence based software is the next era and hence starting now, would add a beneficial skill to my diverse skillset.

3. This might help with future software projects I might work on that require recommender systems to be powered by user sentiments.

4. With this project I explore various NLP techniques, models and algorithms, it is a massive learning opportunity to diversify skills.

Why did I choose this project?

- There are several learning benefits and opportunities in picking sentiment analysis as graduation project such as;

  1. Sentiment analysis is a very relevant field of NLP and hence there loads of previous findings and research pertaining to this field.
  2. Its popularity ensures that there are a lot of resources such as libraries and datasets to build upon.
  3. It can be applied to other fields that use recommender systems and dependence is upon user feedback.
  4. Diversity in interdisciplinary subjects such as statistics, linguistics and computer science henceforth attaining a pretty good-looking skillset.
  5. Contribution to decision making in the movie industry, by means of user feedback.
  6. Innovative software design as Sentiment analysis would improve the overall feel of applications by providing personalised suggestions.

- EXPERIENCE, EXPERIENCE...can't stress this much but sentiment analysis in this movie domain is a very complex field which will give me a hands-on software development practice.

7.2 Future Works

I am not going to keep working on this project, but I plan to be a part of the Artificial Intelligence field in the future.

8. 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.

[6] Detection of Social Network Spam Based on Improved Extreme Learning Machine by Zhijie Zhang, Rui Hou, Jin Yang. Retrieved from https://www.researchgate.net/publication/342226344_Detection_of_Social_Network_Spam_Based_on_Improved_Extreme_Learning_Machine?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19