Optimization of the reviewing process and assessing the popularity of movies

**Sohom Ghosh1\*, Santanu Modak2\*\* and Dr. Abhoy Chand Mondal2\*\*\***

1Department of Computer Science and Engineering, Heritage Institute of Technology, Kolkata – 700107, West Bengal, India

E-mail: [sohom1ghosh@gmail.com](mailto:sohom1ghosh@gmail.com)

2Department of Computer Science, University of Burdwan, Burdwan–713104, West Bengal, India. E-mail: [modaksantanu@gmail.com](mailto:modaksantanu@gmail.com) [abhoy\_mondal@yahoo.co.in](mailto:abhoy_mondal@yahoo.co.in); \*-first author; \*\*-second author; \*\*\*-third author

//improve it if possible Abstract: Opinion analysis has become a flourishing frontier as of late. In this paper, we exhaustively study movie reviews from a popular online database. We randomly sample more than 1000 reviews with titles to train our model. This model is capable of suggesting words during the process of appraising a film. It can intelligently anticipate the words that an appraiser is going to use from the title of his opinion. Furthermore, it has the potential to learn. Whenever it finds that it is unable to suggest words, it learns from the critic’s words. Moreover, this innovative model is able to compute the popularity of a film by examining the opinions. Thus, it improves the task of reviewing by making it quicker and effective. It labels a movie as ‘super-flop’, ’flop’, ‘cool’, ‘hit’ or ‘super-hit’ based on what the reviewers opine. //improve it if possible

Keywords- Opinion mining, sentiment analysis, natural language processing, polarity computation, recommendation system, machine learning, collaborative filtering, product reviews.

1. Introduction//improve it if possible

In this paper we discuss about our unique model to propose words to a reviewer while appraising a film. We use the online database of movies, IMDb to build our training and test set. Firstly, we request the appraiser to enter the name of the movie and the title of his opinion. We analyze this title and predict words he is most likely to use while reviewing. Then, we give him a turn to opine. We examine this opinion and compute score from it. Furthermore, we give him a chance to rate the film. We repeat this process for every user. Finally, for each movie we store every appraiser’s rating. We evaluate the mean from this and declare whether the movie is ‘super-flop’, ’flop’, ‘cool’, ‘hit’ or ‘super-hit’.

This model is beneficial as it makes the process of appraising faster and simpler. Users do not have to spend time wondering for words while reviewing a film. They don’t need to consider about the score they want to assign. They will receive suggestions at each and every step. Moreover, this model has the ability to learn which enhances its efficiency with usage. //improve it if possible

1. Related works (to be written by santanu da)
2. Definitions & notations

* Python/web crawling (if space left, try it and append it here)
  1. Tokenization

The process of splitting a sentence into its constituent words is known as Tokenization. NLTK (Natural Language Toolkit) provides us with Punkt sentence tokenizer. Our model uses split() function to tokenize sentences. After removing trailing blank spaces using strip(), we are splitting the corpus whenever we come across white spaces. Example: Original Sentence: - “The Sky is Blue”; After Tokenization: [‘The’,’sky’,’is’,’blue’]

* 1. POS tagger

POS tagger (or parts of speech tagger) is an inbuilt package of NLTK which maps each word to the parts of speech they belong. Example: Original Sentence: “The opening aerial shots of the prison are a total eye-opener.”

After using POS tagger: - (S The/NNP opening/VBG aerial/JJ shots/NNS of/IN the/DT prison/NN are/VBP a/DT total/JJ eye-opener. /NNP)

* 1. Stemming

Words like ‘behaving’, ‘behave’, ‘behaved’ means the same. They have structural affixes which changes there spelling. To simplify our task we convert these words to a single form i.e. is ‘behave’. This act is referred to as stemming. We use Porter stemmer as a preprocessing step.

* 1. Corpus

Corpus means collection of words. We have built unique corpora separate for positive and negative words. It contains around 6000 words. So, let’s look at a sample from it.

Positive words: - **[':)', ':-)', '=)', 'absolutely', 'adorable', 'accepted', 'acclaimed', 'accomplish', 'accomplishment', 'achievement', 'action', 'active', bountiful', 'bounty', 'brave', 'bravo', 'brilliant', 'bubbly', 'calm', 'celebrated', 'certain', 'champ', 'clean', 'meritorious', 'miraculous', 'motivating', 'moving', 'natural', 'simple', 'skilled', , 'willing', 'wonderful', 'wondrous', 'worthy', 'wow', 'yes', 'yummy', 'zeal', 'zealous', '']**

Negative words:- **[':(', ':-(', '=(', ,** **'abysmal', 'barbed', 'hurtful', 'icky', 'ignore', 'ignorant', 'ill', 'immature', 'imperfect', 'impossible', 'inane', 'inelegant', 'jealous', 'lumpy', 'malicious', 'mean', 'menacing', 'misunderstood', 'moan', 'naive' 'stressful', 'woeful', 'worthless', 'wound', 'yell', 'yucky'']**

1. Polarity

Polarity refers to the positivity or negativity of a word. In this paper, we label positive words as ‘+1’ and negative words as ‘-1’.

1. Mapping \*

Mapping is representation of relations. It refers to a function. For example: *f: X-> Y,* denotes that *f* is a function which maps X to Y. In this paper we use one-to-many and many-to-one mapping. //->pictures, equations, mathematics include if space and time

1. Collaborative filtering \*

For building the recommendation system to suggest words, we use collaborative filtering. It refers to the art of proposing words by gathering interests from the users (collaboration). //->pictures, equations, mathematics include if space and time

1. Mean \*

Mean is a statistical term. It is also referred to as the average. Here, we are finding out the arithmetic mean of some discrete values. The formula for mean is:-

A=\frac{1}{n}\sum_{i=1}^{n} a_i

Here, A denotes the arithmetic mean, ai denotes the discrete values, Ʃ represents summation and n refers to the number of discrete values. //->pictures, equations, mathematics include if space and time

W = \frac{Rv + Cm}{v+m}  Here, *W*= weighted rating, *R*= average for the movie as a number from 0 to 10 (mean) = (Rating), *v*= number of votes for the movie = (votes), *m*= minimum votes required to be listed in the Top 250 (currently 25,000), *C*= the mean vote across the whole report (currently 7.0)

1. Problem definition (// try to improve it)

Given the name of a movie and a title of an opinion about it, we predict the words an appraiser will probably use. We anticipate the score from what he opines. We have built our unique word corpora separate for positive and negative words. We compare the adjectives and adverbs of the opinion with our corpora to detect the degree of polarity of it.

1. Exploring the Model [Work, algorithm, code. Model\*\*]-🡪>>>[algorithm to be written by santanu da from workflow diagram]

Describe the 4 algorithms here.

1. Experimental Evaluation //Try to make the best possible fit out of it

Index: SR-> Suggestive; Rating UR-> User Rating; MP-> Movie Popularity; UN-> User no

Note: Some of the user reviews have been shortened for making this paper compact.

* + - * Case-1

Movie name: Inception

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **UN** | **Title** | **Suggestion** | **User Review** | **SR** | **UR** | **MP** |
| **1** | **Too much....WAY too much** | **average, ordinary, so-so , mediocre , medium** | **What is going on with the IMDb user reviews lately? It's like the masses can no longer be trusted. In the last month, the users have decreed "Airbender" the worst abomination ever, when in fact it's just an average movie.** | **3** | **3** | **cool** |
| **2** | **Insanely Brilliant! Nolan has outdone himself!!** | **really, new, believable, Sure, potential, interesting, cool, special, always, time, good, single, high, better, conventional, high, most, purely, free, creative, highly** | **What is the most resilient parasite? An Idea! Yes, Nolan has created something with his unbelievably, incredibly and god- gifted mind which will blow the minds of the audience away. The world premiere of the movie, directed by Hollywood's most inventive d** | **5** | **5** | **hit** |
| **3** | **Just when I thought I was stuck in a dream (nightmare), I find some realistic reviews of this tosh!** | **Sorry no suggestions for you** | **Perusing the top 250, I was shocked to see this in the list, never mind so near to the top. I make no apology for admitting, that I fell asleep on numerous occasions (though kept re-winding).** | **4** | **3** | **cool** |

* + - * Case-2

Movie name: Apartment 1303 3D

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **UN** | **Title** | **Suggestion** | **User Review** | **SR** | **UR** | **MP** |
| **1** | **Don't waste your money** | **simply, don't, slower, dull, maybe, temporary, slowly, boredom, not, almost, wasn't, never, depressing, aren't, exciting** | **OK, my Summary basically wraps it up. Remember Eddie Murphy's joke about white people moving into haunted houses? "Oh, nice House." "Get out." "just a few ghosts we can handle that.' This movie is an embarrassment.** | **3** | **1** | **Super-flop** |
| **2** | **Truly awful** | **Sorry no suggestions for you** | **Truly awful movie never made it past 15mins, even that was kind. Acting? What acting? You have been warned! I watch and love a lot of horrors, trust me, this shambles was ruined by too much talking to one’s self, too much inane shouting "hello, who's t”** | **2** | **2** | **Super-flop** |
| **3** | **Waste of time!! Mischa Barton can NOT act!!** | **average, ordinary, so-so , mediocre , medium** | **Where to start. This movie was crap. The acting is horrible. That girl cannot act. She sounds mannish and so mono toned throughout the movie.** | **1** | **1** | **Super-flop** |

* + - * Case-3

Movie name: The Shawshank Redemption

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **UN** | **Title** | **Suggestion** | **User Review** | **SR** | **UR** | **MP** |
| **1** | **Tied for the best movie I have ever seen** | **really, new, believable, Sure, potential, interesting, cool, special, always, time, good, single, high, better, conventional, high, most, purely, free, creative, highly** | **The only other movie I have ever seen that affects me as strongly is To Kill a Mockingbird. Both movies leave me feeling cleaner for having watched them.** | **5** | **5** | **Super-hit** |
| **2** | **A classic piece of unforgettable film-making.** | **good, ideal, acting, spectacular, especially, still, entire, little, longer, entire, must-watch** | **Firstly, its setting. The opening aerial shots of the prison are a total eye-opener. This is an amazing piece of architecture, strong and Gothic in design. Immediately, the prison becomes a character. It casts its shadow over most of the film, its tall** | **5** | **4** | **hit** |

* + - * Case-4

Movie name: The godfather

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **UN** | **Title** | **Suggestion** | **User Review** | **SR** | **UR** | **MP** |
| **1** | **"The Godfather" is pretty much flawless, and one of the greatest films ever made** | **good, ideal, acting, spectacular, especially, still, entire, little, longer, entire, must-watch** | **Rather than concentrating on everything that is great about The Godfather, a much easier way for me to judge its quality is on what is bad about it. Almost every film has something that I don't like about it, but I can honestly say that I wouldn't change** | **3** | **4** | **hit** |
| **2** | **Magnificent portrait of organized crime** | **average, ordinary, so-so , mediocre, medium** | **This is by far the best movie ever to give a portrait organized crime; this movie goes deep inside and shows it all inside out.** | **5** | **5** | **hit** |

Here, we can see the popularity of a movie changes. This is because, when each reviewer is rating a movie, the overall score varies. Thus, the popularity of the film is affected.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Movie name | IMDb Rating (out of 10)[=I] | IMDb Rating (out of 5)[=Q=I/2] | Our Rating (out of 5)[=R] | % Accuracy [=A] |
| Inception | 8.8 | 4.4 | 4 | 90.91 |
| Apartment 1303 3D | 2.6 | 1.3 | 2 | 46.15 |
| The Shawshank Redemption | 9.3 | 4.65 | 5 | 92.47 |
| The Godfather | 9.2 | 4.6 | 4 | 86.96 |

A=100-(|Q-R|/Q)\*100

Thus, it is obvious that our model is quite accurate for movies with higher ratings. For poorly rated movies our model seems to be inefficient. But, this is not so. For a movie with lowly rated film the denominator ‘Q’ is comparable to the difference ‘|Q-R|’. This lowers the magnitude of % Accuracy, ‘A’. If we look at the numeric values of Q and R we can easily conclude that they are similar.

1. Conclusions ((//check if something can be added)

This ingenious model is quite handy, fast and fit for use. It saves time. There are few scopes of improvement. We have trained it using about 1000 reviews. It’s advisable to train it with more reviews to enhance its accuracy. Here, we check the polarity of a word by checking its presence in positive or negative corpora. This is not so efficient. Presence of a negative word within 3 words from another negative word makes the sense of sentence positive. Our corpora contain more than six thousand words. It’s better to include more words to it. We used Porter stemmer for pre-processing our data. Porter stemmer sometimes modifies a meaningful word to a meaningless one. It is recommended to use a better stemmer. Instead of suggesting too many words, it is recommended to suggest those with higher frequency of occurrences.

* **References** (some more references to be added by santanu da based on the related works/literature survey he writes)

1. Jacob Perkins, “Python Text Processing with NLTK 2.0 Cookbook”, *PACKT publishing*
2. http://www.enchantedlearning.com/wordlist/positivewords.shtml, http://www.enchantedlearning.com/wordlist/negativewords.shtml; List of positive and negative words.
3. http://computer-ease.com/emotposi.htm, http://computer-ease.com/emotneg.htm List of positive and negative emoticons.
4. <http://www.nltk.org/>
5. Maas, Andrew L. And Daly, Raymond E. And Pham, Peter T. And Huang, Dan and Ng, Andrew Y. And Potts, Christopher. “Learning Word Vectors for Sentiment Analysis” in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 2011.
6. more references to be added by santanu da based on the related works/literature survey he writes

Jobs left: - 1> Header, footer change

2> compact version to 6 pages; single column view

//3> template use

4> Related Works-🡪 Santanu DA

5> Work Flow Diagram to Algorithm -> Santanu DA

6> Exploring the model -> describe the 4 algos as sent by Santanu DA