***Prediction of Box Office Success of Movies Using Hype Analysis of Twitter Data***

Sameer Thigale1, Tushar Prasad2

*MIT College of Engineering,*

*Savitri Bai Phule Pune University, Pune.*

[*sameerthigale08@gmail.com*](mailto:sameerthigale08@gmail.com), [tusshaarprasad@gmail.com](mailto:tusshaarprasad@gmail.com)

**CATEGORY: COMPUTER AND IT PROJECTS**

**PROJECT AREA:** SOCIAL NETWORK ANALYSIS, STATISTICAL PREDICTION, REGRESSION ANALYSIS

**REASONS FOR CHOSING THE PROJECT**

Social media is now everywhere. People create content, and disseminate it to their network. As well they absorb, comment and review thoughts posed by their network.

In recent years, social media has become ubiquitous in our day-to-day life. **Our mindsets are now influenced by what we gather through social networks.** Since social media contains a huge repository of knowledge we decide to investigate in its power to predict real world outcomes.

The topic of movies is of considerable interest among the social media user community. A large number of users are discussing on movies *and have a substantial variance in their opinions*. About Rs. 31,333 Billion (or USD 504 Billion) are added by the Movie industry to the annual GDP. In this case, it becomes important to study the effects of a movie release on the society.

Big budget films are at the risk of making revenue, they need to have an estimate of how much the movie will make and if it is less than expectation what measures the crew should take in order to increase the hype. The production team needs to know the reaction of the mob in social media even before the release of the movie.

**To solve these problems, we have come up with a regression model to accurately predict the box office revenue of movie using Twitter. Our model is based on the simple fact that “A movie well talked about, is well watched”.**

**PROBLEM STATEMENT**

* **To show that a simple regression model built from the rate at which tweets are created about particular movie can outperform market based predictions.**
* **To demonstrate that the amount of attention a movie has strong co-relation to its ranking in future.**
* **To demonstrate how the regression model can be extended to other products of consumer interest.**

**WORKING**

**Steps (in simple terms):**

1. To predict the box office revenue made by the movie, we begin collecting tweets from Twitter Streaming API by running our server 24x7. The tweets are stored in a MySQL database engine. (We have collected more than a million tweets from over 100,000 users.)
2. On the basis of the tweets received, we first verify the assumption made by us. We find that the avg. rate of tweets per hour has a high co-relation with the revenue generated by the movie. We also find a positive co-relation between the revenue generated and category of movie, star-cast, sentiment etc.
3. The regression model equation is as follows:

**I = βa\*A + βp\*P + βd\*D + βs\*S + βc\*C + βe\*E +ε**

**I: Box office revenue**

**A: rate of attention seeking**

**P: polarity of sentiment**

**D: distribution reach**

**S: dummy variable for sequel**

**C: category of movie (Action, romance, thriller, etc)**

**E: Star cast of the movie**

**β: Regression co-efficient vector**

The regression co-efficients are calculated for movies (Here we consider a **“critical period”** for movies as one week before the release of the movie and two weeks after the release of the movie.)

1. We predict the box office revenue for the opening weekend on the release date using the multiple linear regression equation mentioned above.

**CONTRIBUTION TO RESPECTIVE FIELD**

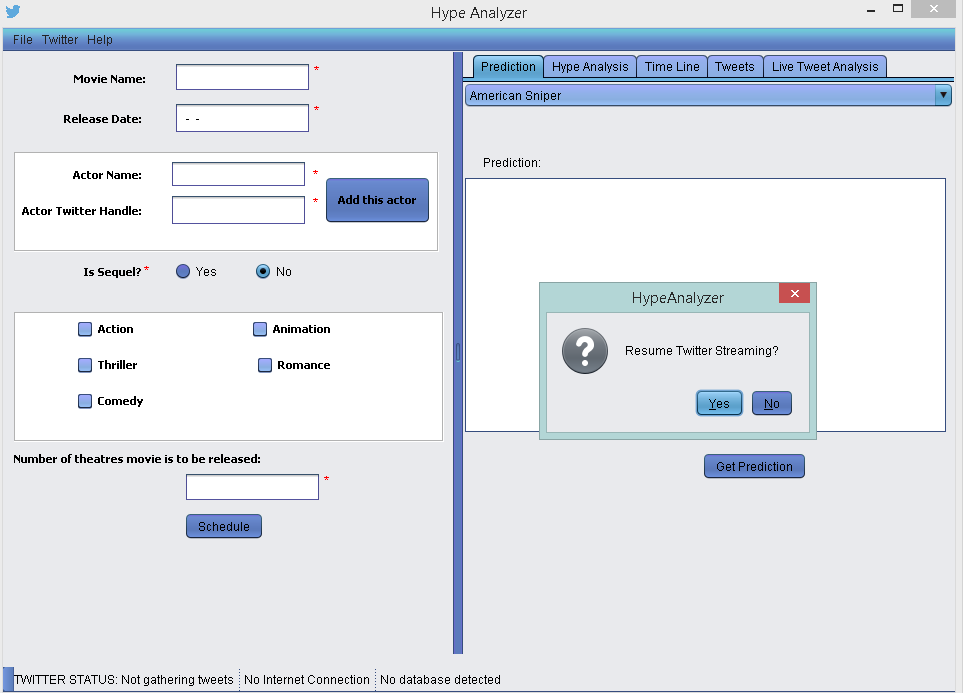
Currently, websites like (Hollywood Stock Exchange) HSX.com are predicting the revenue movies will make, but this prediction is based on the virtual stocks bought by users. None of the websites incorporate the sentiment of users on social media. All of the predictions are either based on polls or opinions, but these kinds of surveys do not provide accurate predictions of movies because either the surveys are on small scale or that entire user opinion is not reflected.

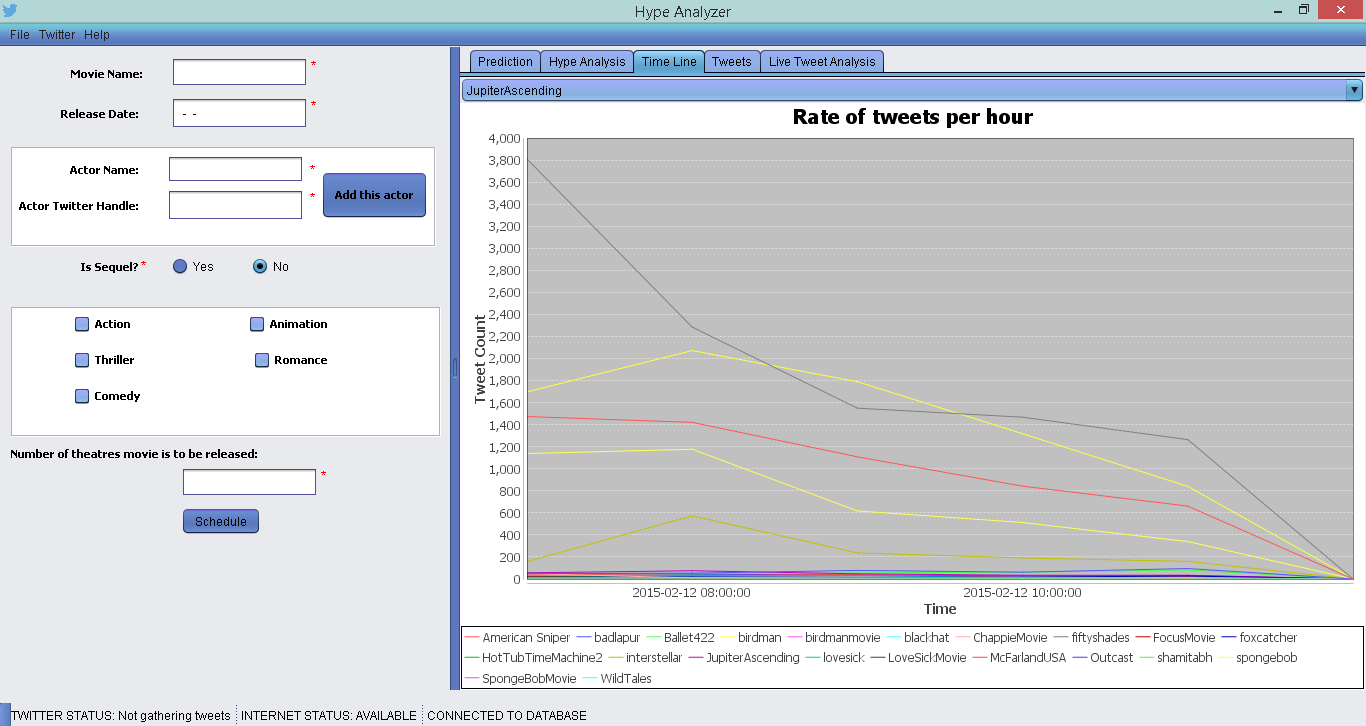
We are incorporating the user opinions on a large scale i.e. not only national but on a global level. This makes our model more accurate than the existing ones. And the result from our model is having a high confidence than the existing ones.

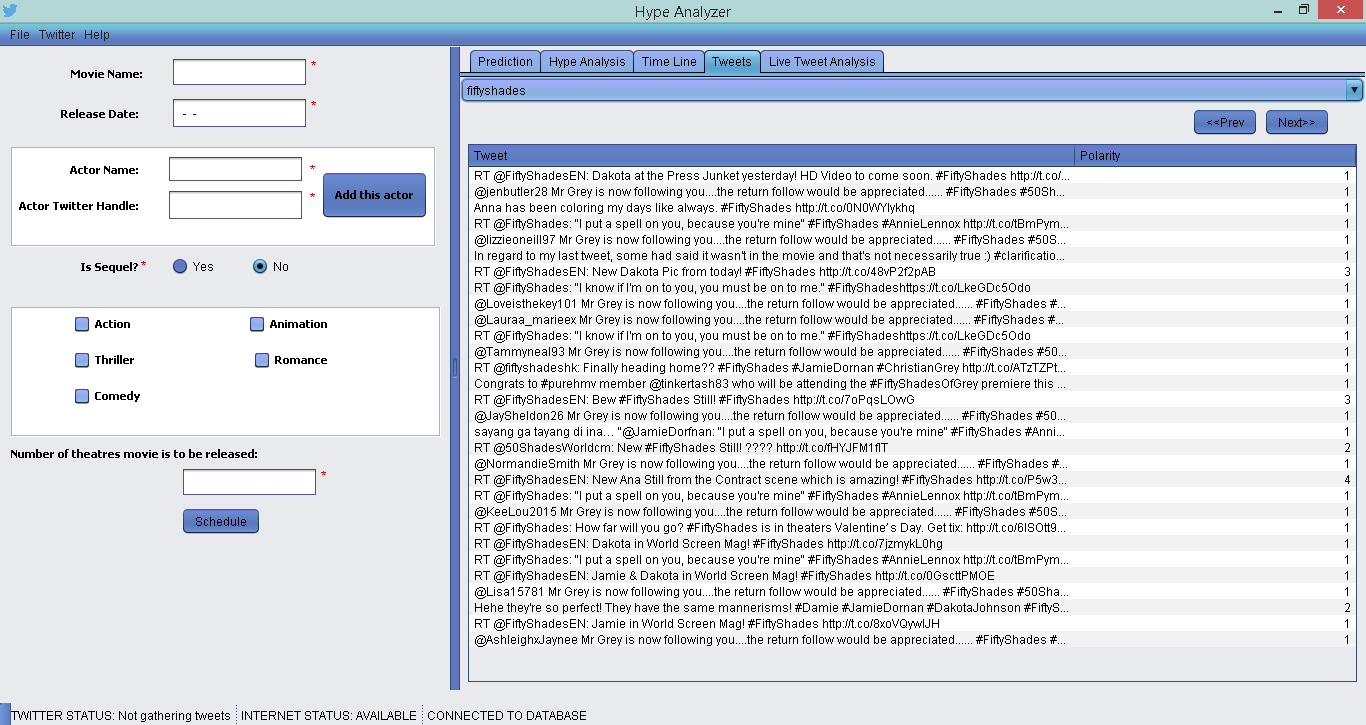
**SALIENT FEATURES**

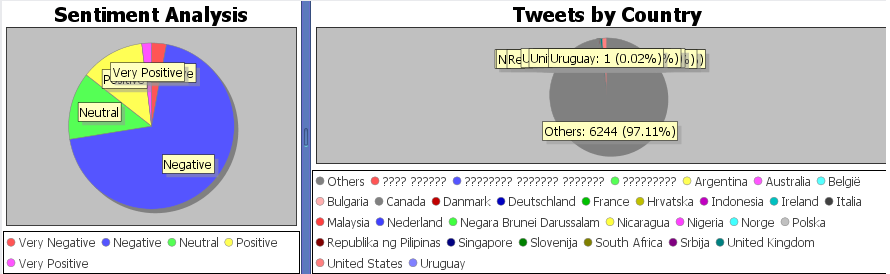
* Prediction of the box office revenue made by movies, given its tweets
* Displays the sentiment expressed by users as Very positive, Positive, Neutral, Negative and Very Negative
* Identification of clusters and frequent used words are shown.
* Project gives insights of the users involved in making tweets.
* It also shows the location from where tweets are coming, the rate of tweets in last six hours, and rate of tweets per second. It also shows the overall user base and its followers.
* The project is completely built on the **“Open source model”** and is built using tools licensed under GNU GPL. It is operating system and architecture independent. That’s why can be run on any PC running Java and MySQL.

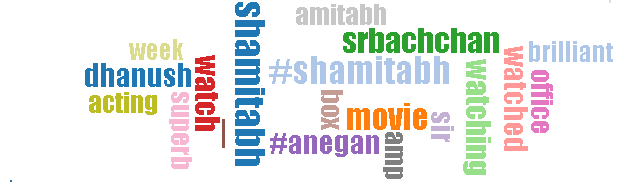
**SCREENSHOTS**

****

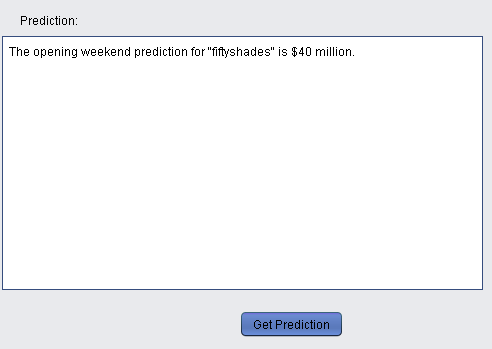
****

****

****

****

****

****

**COMMERCIAL VALIDITY**

* The software can prove extensively useful to the film crew, Production Company, director, producer, production manager to know the investment to output graph.
* Clusters can be identified and appropriate action can be taken in case some clusters have been isolated.
* Theatre owners can schedule their number of shows based on the hype created.
* The software can also act as a repository of tweets, to further data mining and finding patterns.

**PRACTICAL APPLICATION**

Although we have focused on movie revenue prediction here, the method that we advocate can be extended to other products of consumer interest too. This can be done using the same Regression model. Data can be collected from blogs, reviews and other social media too. Collecting the data is important as it measures the rate of chatter effectively.

**FUTURE PLANS**

* In future iterations, the software can incorporate positive effects of **“negative hype”.** For e.g., the effect of the debate on movie PK led to more revenue of the movie
* Incorporation of other factors like music release, effect of songs on success of movies can be incorporated. For e.g., the song “**Tum Hi Ho**” led to the success of the movie Aashiqui 2.
* Sentiments from other social media like Facebook and Google+ can be incorporated.
* Data cleaning process can be automated to increase the accuracy of results.
* The languages accepted by the model can also be extended such that the sentiments expressed in other languages can also be incorporated. This can be done using training of the Stanford NLP tool.

**REFERENCES**

**(BOOKS)**

[1] **FORECASTING- Methods and Applications:** Spyros M., Steven W., Rob H., 3rd Edition, Wiley Publication.

**(RESEARCH PAPERS)**

[2] Sitaram Asur and Bernardo A. Huberman, **Predicting the Future with Social Media**, Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, pp. 492-499, Oct 2010.

[3] Swart William, **Demand Forecasting with Multiple Regression**, IEEE eLearning Library, Dec 2011