# Fuzzy Sentiment Analysis on Microblogs for Movie Revenue Prediction

Niloy Gupta
Computer Science and Engineering
National Institute of Technology
Karnataka, Surathkal, India
niloy_gupta@ieee.org

Abhinav K.R.
Computer Science and Engineering
National Institute of Technology
Karnataka, Surathkal, India
abhinav.kr.90@gmail.com

Annappa
Computer Science and Engineering
National Institute of Technology
Karnataka, Surathkal, India
annappa@ieee.org

*Abstract* – **With the advent of microblogging in recent years, people voice their views about products, especially movies. Microblogs are rich sources of data that can be analyzed to derive useful knowledge like larger public opinion on a product, which can be utilized to derive sales performance patterns. In this paper we propose a novel fuzzy approach for evaluating sentiments expressed in microblogs, which are incorporated in text mining methodologies to predict weekly movie revenues.**

*Keywords* – **Business Intelligence, Fuzzy Logic, Text Mining, Sentiment Analysis, PLSA**

## I. INTRODUCTION

Microblogs offer a convenient platform for the public to express their feelings and opinions about various products, fashion trends, current events and services. With the advent of social media, one person's review can be progressively propagated or rebroadcasted. These features make the microblogging environment a wealth of public opinion data.

Opinion mining is an information retrieval strategy of deriving the opinion or feeling that the text expresses. There has been extensive research in the domain of natural language processing to classify opinions and reviews and estimate the opinion polarity and strength [7]. Lot of studies on web mining, focuses on the relationship between online product reviews and corresponding sales [1][2][3][4]. Since the larger opinion of the public is a clear indication of how the product is favored by the community, it can be used as a yardstick in evaluating the product sales performance. The challenge involves predicting sales or revenue figures, which can be useful for formulating business strategies.

Previous studies in [1][5] focus on linking number of reviews to product performance. While [6] discusses a probabilistic approach in incorporating review sentiments in the prediction process. However, extracting sentiment information from reviews, solely based on word-sentiment frequency is not an effective mechanism. Most text mining techniques do not focus on the pragmatic meaning of a person's review or evaluation. For example, "A brilliant book" carries more sentiment value than "A good book", and thus will have a different effect in the overall sales forecast.

Fuzzy logic provides a powerful tool in quantifying sentiment intensities [9][10][11], but most research is focused towards supervised classification of text reviews as positive reviews and negative reviews; and not towards using the fuzzy approximation to forecast product sales performance.

Movie revenues are freely expressed on the microblogging platforms such as Twitter, Tumblr and Weibo. Aggregated public reviews tend to be more precise in judging the film than expert review. Data mining, natural language processes and machine learning techniques can be employed to extract patterns about public opinions and sentiments [12][13].

In this paper, we present a fuzzy approximation technique to evaluate microblog opinion intensity or degree of polarity of movie reviews. We combine this review intensity with probabilistic latent semantic analysis, to form a regression model that not only encompasses the sentiment quality factor but also the intensity of the sentiment. Results indicate greater accuracy in predicting weekly movie revenue prediction. Section II discusses the process of fuzzifying the reviews collected. It also discusses the preprocessing steps involved. We devise a regression model using the probabilistic topics and fuzzy sentiment intensity values in Section III. We present our results and observations in Section IV. Section V concludes the paper and identifies future scope and applications.

## II. FUZZIFICATION OF SENTIMENTS

We computed the frequency of usage of the adjectives, adverbs and verbs used in the collected microblogs and we eliminated words with low frequencies. This approach saves the system from evaluating all possible combination of words, thereby saving computation time.

Words expressing positive sentiments were assigned a higher fuzzy value whereas words implying a negative connotation were assigned lower fuzzy values. Example of this can be found in Table I. The focus is not on the absolute value but rather on the relative values between the word intensities. To find the overall intensity value of a tweet, we extracted phrases/patterns that expressed sentiments, i.e. which consisted of the sentiment words that were fuzzified in the previous step. The patterns described in Table II were extracted using the POS Tagger[1].

The challenge arises on approximating the overall fuzzy intensity value of the phrase. The following examples illustrate the problem:

[1] *http://nlp.stanford.edu/software/tagger.shtml*

TABLE I. FUZZY INTENSITY VALUES

| Sentiment Word | Intensity Value |
|---|---|
| good | 0.6 |
| amazing | 0.82 |
| horrible | 0.22 |
| bad | 0.4 |
| average | 0.5 |
| very | 0.7 |
| extremely | 0.85 |
| :-) [Happy Smiley] | 0.65 |
| Not,<br>:-( [Sad Smiley],<br>:-P [Sarcasm Smiley] ,<br>:-X [Angry Smiley] | -1 |

TABLE II. PATTERNS OF EXTRACTED PHRASES

| Pattern | Word 1 | Word 2 | Word 3 | Word 4 |
|---|---|---|---|---|
| 1 | Adjective/Verb | Noun | - | - |
| 2 | Adverb | Adjective/Verb | Noun | - |
| 3 | Adverb | Adverb | Adjective/ Verb | Noun |
| 4 | Negation | Pattern 1/2/3 | - | - |
| 5 | Pattern 1/2/3/4 | Emoticon | - | - |

1. "XYZ was an awesome movie"
2. "XYZ was a bad movie"
3. "XYZ was not an awesome movie"
4. "XYZ was a very good movie"
5. "XYZ was not a very good movie"
6. "XYZ was a funny movie :D "
7. "XYZ was super awesome :P "

The above examples illustrate the complexity involved in identifying the tweet polarity as microblog statements do not conform to the strict rules of grammar, involve slang, use of emoticons and web jargons. Thus standard algorithms that use dictionaries cannot be utilized in classifying opinions that are expressed on microblogs.

We divide the classification problem into four cases and formulate a fuzzyfication technique for each case. The fuzzyfication function is illustrated in Fig 1.

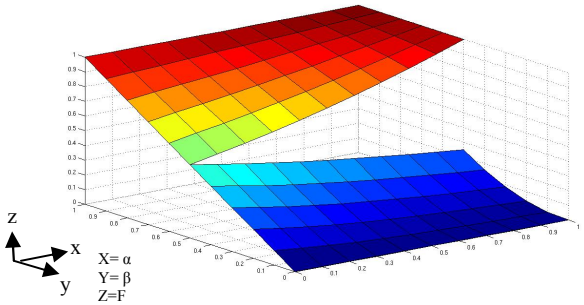


Fig 1. Fuzzification Function

Let us define the functions F( ) and f( ) as follows.

F - Final fuzzy intensity value of the phrase P

f - Fuzzy intensity value of the sentiment word

*Case 1*: Pattern1 (Adjective/Verb, Noun) : $\rho_1(\alpha,\beta)$

$$F(\rho_1) = f(\alpha)$$

TABLE III. FUZZIFICATION EXAMPLE CASE 1

| Phrase | Fuzzy Intensity Value |
|---|---|
| good movie | 0.6 |
| Horrible movie | 0.22 |

*Case 2:*

Pattern2 (Adverb,Adjective/Verb, Noun) : $\rho_2(\alpha,\beta,\gamma)$

$$F(P2\ (\alpha,\beta,\gamma)) = \{\ F(\rho_1(\beta))^{1+f(\alpha)}\ ,\ F(\rho_1(\beta))<0.5$$
$$F(\rho_1(\beta))^{1-f(\alpha)}\ ,\ F(\rho_1(\beta))>=0.5\ \}$$

TABLE IV FUZZIFICATION EXAMPLE CASE 2

| Phrase | Fuzzy Intensity Value |
|---|---|
| Very good movie | 0.858 |
| Very bad movie | 0.129 |

*Case 3:*

Pattern3 (Adverb, Adverb, Adjective/Verb, Noun) : $\rho_3(\alpha,\beta,\gamma,\mu)$

$$F(\rho_3(\alpha,\beta,\gamma,\mu)) = F(\alpha,F(\rho_2(\beta,\gamma,\mu),\ \mu)$$

TABLE V. FUZZIFICATION EXAMPLE CASE 3

| Phrase | Fuzzy Intensity Value |
|---|---|
| Extremely very good movie | 0.97 |

*Case 4:* Pattern4 (Negation, Pattern 1/2/3) : $\rho_4(\alpha, \rho_i)$

where i=1,2,3

$F(\rho_4(\alpha, \rho_i)) = 1 - F(\rho_i)$

TABLE VI. FUZZIFICATION EXAMPLE CASE 4

| Phrase | Fuzzy Intensity Value |
|---|---|
| Not very good movie | 0.142 |

*Case 5:* Pattern1/2/3/4 + Emoticon : $\rho_5(\rho_i, \alpha)$

where i=1,2,3

If α is a negation, $F(\rho_5(\rho_i, \alpha)) = 1- F(\rho_4(\alpha, \rho_i))$

Else, $F(\rho_5(\rho_i, \alpha)) = F(\alpha, F(\rho_i))$

TABLE VII. FUZZIFICATION EXAMPLE CASE 5

| Phrase | Fuzzy Intensity Value |
|---|---|
| amazing movie :-X | 0.18 |
| Amazing movie :-) | 0.93 |

The above recursive fuzzyfication function will consider the cases of overlapping patterns such that the fuzzy intensity value is calculated only for the parent pattern, not its subsets.

The total fuzzy intensity value of all the tweets collected for a movie is the average of all the fuzzy intensity values of all the phrases/patterns in the tweets.

$\psi(M)=\Sigma_{i=0}^{n} F(\rho_i)/n$,

Where n – Total number of phrases

M – Movie for which the tweets are collected

$\psi$(M)- fuzzy sentiment factor for the particular movie M.

## III. REGRESSION MODELING

Regression can be used to model the relationship between weekly box office revenue and its determining parameters. Yu and Liu in [6] use the traditional probabilistic latent semantic analysis (PLSA) [7] model to identify hidden factors that are embedded in movie reviews. Instead of using the traditional "vanilla bag of words" they use only appraisal words for sentiment classification [8].

We use the same PLSA technique of identifying hidden sentiments, but since our focus is on microblogs, mere use of traditional appraisal words will not provide useful hidden sentiments in an informal microblogging environment.

Statements such as "Movie XYZ <3 <3" conveys string sentiments, but an algorithm based on appraisal dictionaries cannot capture the hidden sentiment factors. To carry out this technique, we generate a set of sentiment words or symbols, based on processing pre-collected data using semi-automatic techniques. These keywords are passed to the PLSA algorithm which generates probabilistic topics.

Weekly box office movie collections are strongly dependent on the weekend collection. This temporal relationship can be exploited to develop a strong regression model, which can predict with greater accuracy. One might suggest to the daily collections of all the preceding days to develop a more accurate regression model, but since our target is to develop an intelligent system that can make long term predictions based on minimal data, the use of all daily collections is unwarranted. Also in most film industries, accurate daily collections are not as easily available as with opening weekend collections. Thus, one must also consider the pragmatic aspects such as availability of data on movie collections while developing such a system.

Based on experimentation, we develop the following regression model

$$log(\Phi_P/ \Phi_W)= \alpha_1\tau_1 + \alpha_2\tau_2 + \alpha_3\tau_3 + \alpha_4\psi(M)$$

Where,
$\Phi_P$ = Predicted Weekly Movie Revenue
$\Phi_W$ = Collected Weekend Movie Revenue
$\alpha_x$ = Parameters predicted by the Regressive Model
$\tau_x$ = Sentiment Topics generated by PLSA
$\psi(M)$ = Fuzzy Sentiment Factor

This regression model has been customized for the domain of movie and box office revenues. The model can be modified or extended for other products as well.

## IV. EXPERIMENTAL SETUP AND RESULTS

In our study, we use the Twitter platform to collect tweets/movie reviews using the Twitter API to build the training set. Data about daily movie revenues was collected from IMDB and BoxOfficeMojo. Also, we used Part-Of-Speech(POS) Tagger to classify each word in the tweet into its corresponding part of speech. This was useful in extracting sentiment features from the tweets.

A training set was created using the twitter results by querying movie names and tags.

Initially, all the tweets were POS Tagged and phrases were extracted using pattern matching tools. The overall sentiment value was calculated using the fuzzy system developed.

The process of creating a dictionary of sentiment and appraisal words, terms and emoticons involved analyzing the tweets and extracting the ones that were used extensively. It was observed that these frequently used key-words formed less than 10% that of traditional dictionary appraisal words. This is clear indication that a small set of words form a determining factor for sentiment analysis. This method also takes care of the change in online opinion expression with time. Thus, the dictionary that we develop will evolve to changes in the language used by web users.

We collected the data for the movies released between June 2011 and March 2012 for our experimentation purposes. The regression model developed was tested on various movies and the results obtained indicate an accuracy of 90–95%. These results prove to be very encouraging considering the fact that most of the data was derived from online expression of opinion. One might observe that the use of online sentiment expression for predicting box office revenue provides a convenient method rather than considering other salient parameters such as expert ratings, movie budget, cost of publicity, regional statistics, etc. Thus the use of micro blogs overcomes the challenge of collecting complex statistics about these movie parameters. Also, the choice of fuzzyfication of sentiments provides better results as compared to predictive systems that do not consider the quality of the reviews as in [6]. Merely using user ratings as a quality factor does not capture the public sentiments. As a numerical value on the number scale, given by a customer of a product might not completely express his opinion. But fuzzyfication provides the tool of converting sentiment expressed in words, into quantifiable parameters that can effectively express user sentiment.

This model takes into consideration various anomalies such as a movie with tremendous publicity and budget might not be received well at the box office. While movies, that do not have media publicity or brand value might perform well, because of a positive public opinion. Since, the success or failure of a movie ultimately depends upon the people's choice, which is openly expressed on microblogging platforms; thus using tweets allows us to formulate a highly accurate predictive system.

## CONCLUSION

We presented a fuzzy sentiment analysis approach that can be complemented with existing web mining and machine learning techniques predict product sales

performance. In our work, we considered the case of movies, but the idea and approach can be extended to other products as well, for which the customer opinion is available online and in abundance.

From our experimental results we can conclude that our proposed methodology provides an effective prediction system, based on which calculated business decisions could be formulated.

Especially in the today's world of Web 2.0; where public sentiments are freely expressed and available fuzzy approximation techniques combined with natural language processing and learning algorithms can provide us with unique results to help understand the public mood and preference.

## REFERENCES


[1] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The predictive power of online chatter," in KDD, 2005, pp. 78–87.

[2] A. Ghose and P. G. Ipeirotis, "Designing novel review ranking systems: predicting the usefulness and impact of reviews," in ICEC, 2007, pp. 303–310.

[3] Y. Liu, X. Huang, A. An, and X. Yu, "ARSA: a sentiment-aware model for predicting sales performance using blogs," in SIGIR, 2007, pp. 607–614.

[4] Y. Liu, X. Yu, X. Huang, and A. An, "Blog data mining: The predictive power of sentiments," in Data Mining for Business Applications. Springer, 2009, pp. 183–195..

[5] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in WWW, 2004, pp. 491–501.

[6] Xiaohui Yu; Yang Liu; Xiangji Huang; Aijun An; , "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain," Knowledge and Data Engineering, IEEE Transactions on , vol.24, no.4, pp.720-734, April 2012doi: 10.1109/TKDE.2010.269

[7] T. Hofmann, "Probabilistic latent semantic analysis," in UAI, 1999

[8] C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal groups for sentiment analysis," in CIKM, 2005, pp. 625–631.

[9] Nadali, S.; Murad, M.A.A.; Kadir, R.A.; , "Sentiment classification of customer reviews based on fuzzy logic," Information Technology (ITSim), 2010 International Symposium in , vol.2, no., pp.1037-1044, 15-17 June 2010

[10] Animesh Kar, Deba Prasad Mandal, Finding Opinion Strength Using Fuzzy Logic on Web Reviews, International Journal of Engineering and Industries, volume 2, Number 1, March, 2011

[11] Subasic, P.; Huettner, A.; , "Affect analysis of text using fuzzy semantic typing," Fuzzy Systems, 2000. FUZZ IEEE 2000. The Ninth IEEE International Conference on , vol.2, no., pp.647-652 vol.2, 2000

[12] Brendan O'Connor, Ramnath Balasubramanyan, From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series, Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, May 2010

[13] Antweiler, W., and Frank, M. Z. 2004. Is all that talk just noise? the information content of internet stock message boards. Journal of Finance 59(3):1259–1294.