Are Emotions Associated with Social News Popularity?

Asra Sakeen Wani
asraw@iiitd.ac.in
Indraprastha Institute of Information Technology Delhi

Divyanshu Kumar Singh
divyanshu17048@iiitd.ac.in
Indraprastha Institute of Information Technology Delhi

Prachi Arora
prachi17075@iiitd.ac.in
Indraprastha Institute of Information Technology Delhi

Pramil Panjawani
pramilp@iiitd.ac.in
Indraprastha Institute of Information Technology Delhi

ABSTRACT

News social popularity plays a key role for any news outlets as it help them to predict and tune their headlines and content to enhance the user interaction. In this paper, we are extending the work proposed by Kumar et al., 2019 [10]. We can predict popularity for articles, images or any other online content, which can affect market strategies, recommendation systems, and even policy-making. We try to classify news as popular or unpopular using early view count and also perform emotion analysis on various data-sets. We attempt to implement various Machine learning algorithms for the above tasks and evaluate using metrics from the original work. We also aim to look from a physiological perspective and model above using valence and emotions. Using data-sets from various news organizations and aim to build a generalized model that can be used for various applications. Up till now, our focus was on replicating the results of the baseline paper. It has been observed that for, category relevance feature, our model outperforms the the baseline model. However, the difference can be attributed to several reasons out of which, one is the usage of a different set of word embeddings.

KEYWORDS

Social Media, News, Emotional Valence, Machine Learning

ACM Reference Format:
Asra Sakeen Wani, Divyanshu Kumar Singh, Prachi Arora, and Pramil Panjawani. 2020. Are Emotions Associated with Social News Popularity?. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Information sharing has evolved over the decade, starting from the letter writing to writing email to communicate. Social media is one such platform which has leveraged the ICT(Information and Communication Technology) e.g Twitter, Facebook, WhatsApp and etc. These technologies have not only helped personal communication but the mass communication is also enhanced for e.g News sharing. Every major news outlets use social media for instant information sharing, hence it could be said that information sharing via News outlets have gained the shape of marketing. A study done by The Pew Research Center in the US, reported that 68% of adult American used Facebook and 74% of them used it everyday[18]. Therefore, it is essential for the news outlets to identify the potential platform and user, and to try reach out to the audience in an effective manner. A lot articles are published everyday in various news articles and they simply push the articles onto their respective social media pages without analysing how the articles would be perceived via the general audience. Hence, the notion of News popularity comes into the picture. News popularity have been studied by various researchers[4][7], every research finds a key metric to judge the population be it page views[1]; most comments[3] or the most shares.

In general, the notion of predicting news popularity is very difficult because the way news is presented varies from different sources. These sources are also dependent on the very nature of socio-cultural dynamics which could be clearly seen in the regional news sources. This paper builds onto the existing work done by Kumar et al., 2019 [10], their major research question revolved around generalising such models and also introducing the notion of emotion from the news headlines data-set that they collected for the period of 10 consecutive months. Their analysis has three major contributions : a.) their model takes into account the correlation between the news popularity; b.) their key features does not rely on the data/time of the post and retweets after publication and c.) development of large and novel data-set.

The results were derived using two approaches a.) Correlation analysis for very emotional valence feature and b.) Support Vector Machines for popularity prediction. For the correlation analysis they measure the emotional intensity scores with association to social popularity and performed a two-sample t-test. Insights from this indicated a clear pattern as the all the five emotional intensity (i.e Valence, Joy, Anger, Fear and Sadness) were found to be associated with the news social popularity. For examples, top headlines had significantly higher intensities of valence and joy, but lower anger, fear and sadness. For the social popularity prediction, they performed a 10-fold cross validation for both the classification and regression task.

2 LITERATURE REVIEW

In this section we situate the literature review carried out in the same direction of work. The authors in [12] paper aimed at predicting the news articles popularity depending on the on Mashable dataset. The dataset comprises of 58 heterogeneous features of

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

news articles by Mashable which is a well known digital media site that normally shares web journals, articles and short articles. The popularity of a new article is predicted based on the factors such as time, length, polarity, data channel. The authors in the paper have implemented eleven models out of which Gradient boost algorithm outperformed by achieving an accuracy of 79.9% on the Online News Popularity dataset. On the similar lines of work by Yaser et al, [11] the paper focuses on predicting the popularity of the new article based with an objective of integrating multiple popularity measurements across various news directs. The work carried out primarily targets on predicting the news popularity in a local context and utilize the number of page views of an article as a surrogate for its popularity. In this study the problem of predicting new popularity is casted as a regression problem and the features are extracted from the news articles up to 30 minutes once the news is posted, and the based on these features predicting the number of page views an article will receive in 24 hours. The best model of the study is deployed in the real-time system in The Washington Post. The experiments in the study use 10 fold cross validation and states the performance of the models based on the average adjusted R2 values.

Some of the papers study early reaction to an article and try to predict the popularity using these models [2][6][13][14][13].We know that there is some correlation in early and late popularity and number of views, likes/ dislikes or any voting mechanism, we still don't know what drives them to success [8][19].Salganik et al.[17] conducted a major study that addressed this question experimentally by measuring the impact of content quality and social influence on the eventual popularity or success of cultural artifacts.

The authors in [15] collected a dataset by scraping news articles from Digg's website on the technology section in May and June 2006. The model incorporates parameters such as early votes and website design.They developed a stochastical model framework relating users' individual choices to their aggregate behavior, which is, in turn, related to the changes in the state of a single story. As a story is promoted, it shows up at the top of the front page list. If a story does not reach this threshold within a day after submission it gets removed. They developed a simple model, which did not consider the number of fans for a story's voters, has a lower correlation, 75%, with the observed numbers and a larger RMS error for stories. They developed a model of social voting on Digg, which explains how the number of votes received by a story changes with time. Knowing how interesting a story is and how connected the submitter fully shows the increase in the number of votes the story receives. This leads to an insight that a model can be used to predict the story's popularity from the initial reaction of users to it.

The authors in [5], take a psychological approach to understanding what makes a news article popular. Data set was collected from the New York Times articles published over a three-month period, the authors examine how emotion affects the virality of the content. The results show that positive content is usually more viral than negative content, but the relationship between emotion and popularity is more difficult to study than just based on valence alone. Virality is partially driven by physiological arousal. High-arousal positive (awe) or negative (anger or anxiety) emotion-inducing content is usually more viral than the content that evokes low-arousal emotions (e.g., sadness).We also see that awe, a high-arousal positive emotion, is also linked to virality.

3 BASELINE PAPER

This section briefly explains the baseline paper[10]. Sentiments analysis, social popularity, fake news are all being studied ubiquitously in the domain of natural language processing and machine learning. Also, with advancement of deep learning, the research has been accelerated to new heights. In this works, the author are attempt to study the possible cor-relational understanding between the social news popularity and the emotional salience of those posts. The authors created a novel dataset consisting of 47,611 English news headlines from six major publisher in Singapore. As a first step towards understanding the importance of emotional salience as compared to other features, we decided to replicate some of the predictive results in the this paper.

3.1 Dataset and Code

The dataset for the baseline paper has been received from the authors. The dataset contains news headlines with title, summary, post message etc and related metadata corpora (timestamps, news category etc.) which is collected from six English news publishers in Singapore - (The Straits Times, Today Online, Channel NewsAsia, Mothership, The Online Citizen, The Independent Singapore), labelled with shares and likes. Dataset and code both can be found here[1].

3.2 Architecture

3.3 Feature Selection and Results

Up till now, we have worked on two out of the five features extracted in the baseline paper.

- *Relevance-News Category Features:* As highlighted in the baseline paper, there exists a direct correlation between news category and the readership an article amasses. Now to replicate the results, the baseline does not clarify the methods used. The only knowledge that we had at our hand was, the use of libSVM for both regression and classification tasks. To add correlation between the headlines, news summary and the category, we decided to train a word embedding model(Fasttext.ai) trained on the summary. Each sentence was tokenized into words for which a word embedding was obtained. Fasttext was mainly used because, since it breaks individual word into an n-gram, it can assign a word embedding to unseen words like in case of some proper nouns. The training set for the SVM consisted of averaged word embedding vector for all the words in a sentence, concatenated with the word embedding of the assigned category of the article. The results have been summarised in Table 1 and Table 2. The slight variation in the results can be attributed to usage of different word embeddings, and the fact that there were 43 categories that were found in the dataset as compared to 40, which was actually stated in the paper. Overall, there was a bigger discrepancy in the regression model as compared to classification task.

[1] https://github.com/sdiv0/MCA-Project-Group-2

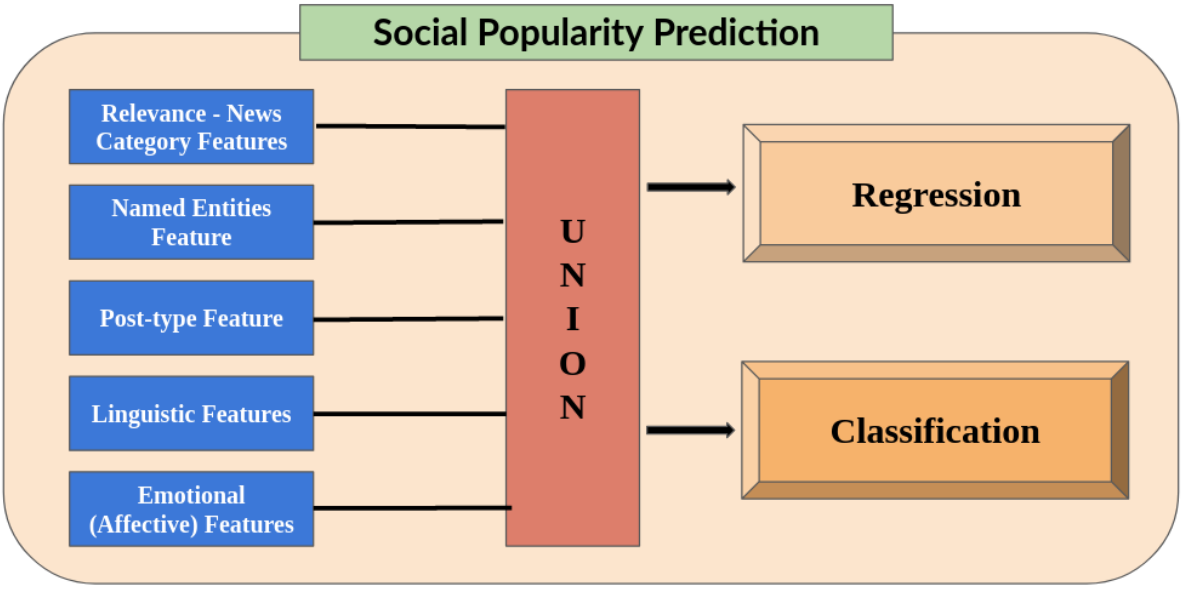

Figure 1: Social Popularity Prediction

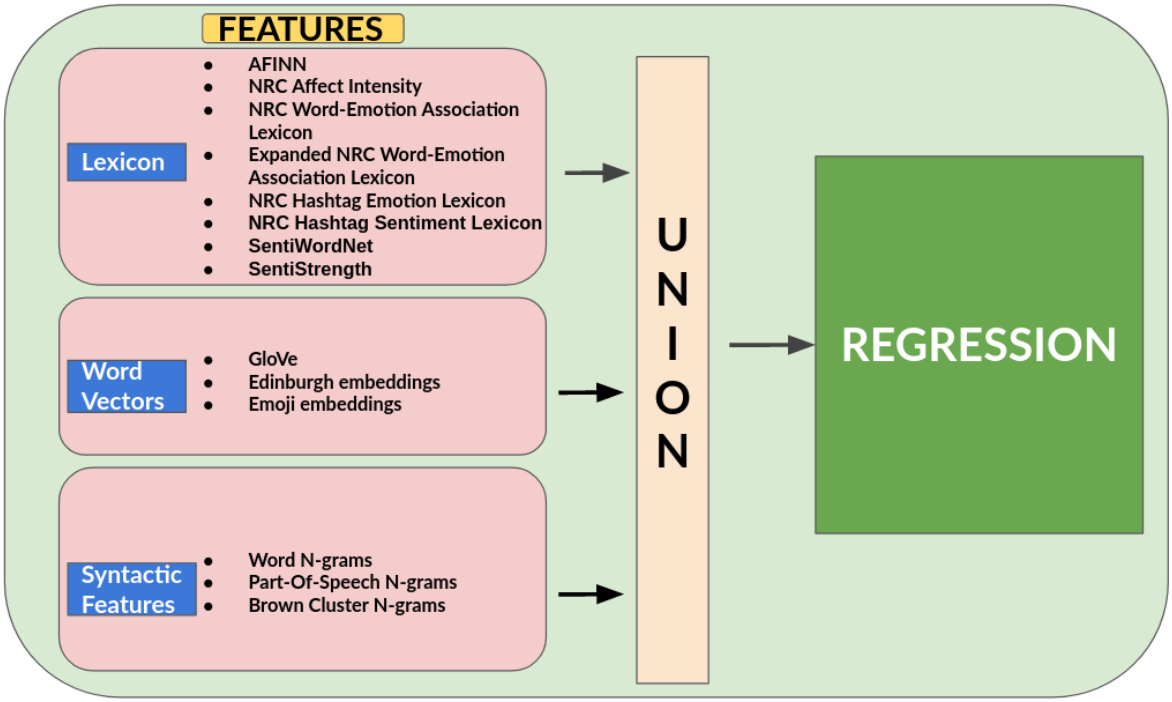

Figure 2: EmoInt : Emotional Salience Model

- *Prominence and Proximity Features - Named Entity Features* - As implemented in the original paper, we extracted the names of people, locations and organizations from the headlines using Stanford NER (Named Entity Recognizer)[2]. The dataset used by the authors to categorise the local locations and organizations were unavailable, hence, this variable was ignored. The entities so obtained were vectorized and were used for prediction. In the results obtained for classification, the recall is significantly higher than that in the original

[2]https://nlp.stanford.edu/software/CRF-NER.html

Model	R-Square	MAE	Kendall's Tau
Original Paper	0.039	1.124	0.106
Our Approach	**0.118**	**1.449**	**0.245**

Table 1: Relevance-News Category Features: Regression

Model	Precision (P)	Recall (R)	F1	AUC
Original Paper	0.118	0.433	0.178	0.553
Our Approach	0.095	**0.436**	0.156	**0.609**

Table 2: Relevance-News Category Features: Classification

Model	R-Square	MAE	Kendall's Tau
Original Paper	0.039	1.125	0.143
Our Approach	-0.0136	0.706	0.015

Table 3: Named Entity Features: Regression

Model	Precision (P)	Recall (R)	F1	AUC
Original Paper	0.129	0.515	0.205	0.568
Our Approach	**0.535**	0.316	**0.397**	0.518

Table 4: Named Entity Features: Classification

Model	R-Square	MAE	Kendall's Tau
Original Paper	0.035	1.135	0.140
Our Approach	**0.125**	**1.449**	**0.240**

Table 5: Richnesss-Post Type Features: Regression

Model	Precision (P)	Recall (R)	F1	AUC
Original Paper	0.185	0.465	0.215	0.557
Our Approach	**0.546**	**0.498**	**0.521**	**0.608**

Table 6: Richness-Post Type Features: Classification

Model	R-Square	MAE	Kendall's Tau
Original Paper	0.061	1.115	0.154
Our Approach	**0.081**	**0.0692**	**0.202**

Table 7: Linguistics Style Feature: Regression

Model	Precision (P)	Recall (R)	F1	AUC
Original Paper	0.128	0.519	0.205	0.595
Our Approach	**0.507**	0.320	**0.392**	0.527

Table 8: Linguistics Style Feature: Classification

paper. The difference between other parameters is of lesser margin and insignificant as compared to the recall. We overall obtained a poorer performance in the regression task. One reason for that might be the usage of only top 10000 headlines because of lack of availability of strong GPUs to perform entity recognition on the entire dataset.

- *Richness-Post Type Featuresm* - Here, we simply took the same approach as in the case of category features. The results obtained for classification and regression by our method are overall better than that in the original paper.
- *Linguistics Style Features* - A simple six dimensional vector was extracted from word-level linguistics features that are associated with common headline designs. These include headline length, total words, and words with more than six letters (i.e., complex words), uses of upper case, superlative and comparative words in the headlines. Here, to get comparative and superlative terms in the headlines, we used the Stanford POSTagger[3]. The overall performance was better in case of regression. Here, this too can be attributed to the consideration of top 20 percent of headlines with the largest like counts.
- *Emotional and Affective Features* - In general to predict the emotional salience, the study uses some sort of annotation from the user to actually gaze how the human perceive the emotions. This work though, use a different approach for the classification of the items into different emotions and i.e Best Worst Scaling. This was introduced by Mohammad and Bravo-Marquez [16] to reduce the subject bias of likert scales. Instead of asking annotators to rate in a scale of the intensity of a specific emotion (e.g., anger) they see from a tweet, they asked annotators to rank the best and worst examples of the intensity of emotions among n text examples.
  To analyse the emotional intensity, the authors used 5 independent SVM-based algorithms, named as CrystalFeel [9]. The algorithms were trained on the annotated tweets data from [16]. The collection of algorithm, process each text(e.g headlines, tweets and etc.) and returns the fives dimension of the emotional intensity scores.
  However, due to unavailibity of CrytalFeel, and unclear description of all the 28 features used in the paper, we resorted to use EmoInt[4], which has 14 feature extractors clubbed into three categories:
  – Lexicon Features
  – Word Vectors
  – Syntax features
  All our results obtained from our experiments are summarised in these tables.

4 CONCLUSION

The paper explores the emotional features and looks build up an emotional salience based model for predicting and understanding the social popularity of news. Despite the unavailibity of clear documentation of the CrytalFeel used in the baseline line paper, and unclear description of all the features used in the paper, we resorted to use EmoInt which has 14 feature extractors clubbed into three categories - Lexicon features, Word Vectors, Syntax features. Our approach and results performed better in some cases from the baseline paper as is detailed in the tables above.

[3]https://nlp.stanford.edu/software/tagger.shtml

[4]https://github.com/SEERNET/EmoInt

Model	Precision (P)	Recall (R)	F1	AUC
Original Paper	0.185	0.465	0.215	0.557
Our Approach	**0.546**	**0.498**	**0.521**	**0.608**

Table 9: Richness-Post Type Features: Classification

Model	R-Square	MAE	Kendall's Tau
Original Paper	**0.178**	1.037	**0.281**
Our Approach	0.0355	**0.710**	0.130

Table 10: All Feature Except Emotional Features: Regression

Model	Precision (P)	Recall (R)	F1	AUC
Original Paper	0.241	**0.519**	0.247	**0.683**
Our Approach	**0.582**	0.194	**0.291**	0.531

Table 11: All Features Except Emotional Feature: Classification

Model	R-Square	MAE	Kendall's Tau
Original Paper	**0.074**	**1.109**	**0.169**
Our Approach	0.0177	1.554	0.107

Table 12: Emotional Salience(EmoInt) Feature: Regression

Model	Precision (P)	Recall (R)	F1	AUC
Original Paper	.154	**0.475**	**0.232**	**0.635**
Our Approach	**0.583**	0.0918	0.158	0.515

Table 13: Emotional Salience(EmoInt) Feature: Classification

Model	R-Square	MAE	Kendall's Tau
Original Paper	0.211	1.014	0.306
Our Approach	**0.937**	**0.0486**	**0.982**

Table 14: All Features (Including EmoInt Features): Regression

Model	Precision (P)	Recall (R)	F1	AUC
Original Paper	0.254	0.321	0.284	0.716
Our Approach	**0.535**	**0.473**	**0.502**	0.5431

Table 15: All Features (Including EmoInt Features): Classification

5 ACKNOWLEDGEMENT

We would like to thank Prof. Rajiv Ratan Shah for his constant support and encouragement throughout the project.

6 INDIVIDUAL CONTRIBUTION

- Asra - Literature Review, Machine Learning - 25%
- Prachi - Natural Language Processing and Machine Learning - 25%
- Pramil - Affective Computing - 25%
- Divyanshu - Affective Computing - 25%

REFERENCES

[1] Sofiane Abbar, Carlos Castillo, and Antonio Sanfilippo. 2018. To post or not to post: Using online trends to predict popularity of offline content. In *Proceedings of the 29th on Hypertext and Social Media*. 215–219.

[2] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S Yu. 2008. Identifying the influential bloggers in a community. In *Proceedings of the 2008 international conference on web search and data mining*. 207–218.

[3] Carl Ambroselli, Julian Risch, Ralf Krestel, and Andreas Loos. 2018. Prediction for the Newsroom: Which Articles Will Get the Most Comments?. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. 193–199.

[4] Roja Bandari, Sitaram Asur, and Bernardo A Huberman. 2012. The pulse of news in social media: Forecasting popularity. In *Sixth International AAAI Conference on Weblogs and Social Media*.

[5] Jonah Berger and Katherine L Milkman. 2012. What makes online content viral? *Journal of marketing research* 49, 2 (2012), 192–205.

[6] Riley Crane, Didier Sornette, et al. 2008. Viral, Quality, and Junk Videos on YouTube: Separating Content from Noise in an Information-Rich Environment.. In *AAAI Spring Symposium: Social Information Processing*. 18–20.

[7] Nicholas Diakopoulos and Arkaitz Zubiaga. 2014. Newsworthiness and network gatekeeping on twitter: The role of social deviance. In *Eighth International AAAI Conference on Weblogs and Social Media*.

[8] Vicenç Gómez, Andreas Kaltenbrunner, and Vicente López. 2008. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th international conference on World Wide Web*. 645–654.

[9] Raj Kumar Gupta and Yinping Yang. 2018. CrystalFeel at SemEval-2018 Task 1: Understanding and Detecting Emotion Intensity using Affective Lexicons. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana, 256–263. https://doi.org/10.18653/v1/S18-1038

[10] Raj Kumar Gupta and Yinping Yang. 2019. Predicting and Understanding News Social Popularity with Emotional Salience Features. In *Proceedings of the 27th ACM International Conference on Multimedia*. 139–147.

[11] Yaser Keneshloo, Shuguang Wang, Eui-Hong Han, and Naren Ramakrishnan. 2016. Predicting the popularity of news articles. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 441–449.

[12] Aasim Khan, Gautam Worah, Mehul Kothari, Yogesh H Jadhav, and Anant V Nimkar. 2018. News Popularity Prediction with Ensemble Methods of Classification. In *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 1–6.

[13] Kristina Lerman. 2006. Social networks and social information filtering on digg. *arXiv preprint cs/0612046* (2006).

[14] Kristina Lerman and Aram Galstyan. 2008. Analysis of social voting patterns on digg. In *Proceedings of the first workshop on Online social networks*. 7–12.

[15] Kristina Lerman and Tad Hogg. 2010. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th international conference on World wide web*. 621–630.

[16] Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion Intensities in Tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*. Association for Computational Linguistics, Vancouver, Canada, 65–77. https://doi.org/10.18653/v1/S17-1007

[17] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science* 311, 5762 (2006), 854–856.

[18] A Smith and M Anderson. 2018. A majority of Americans use Facebook and YouTube, but young adults are especially heavy users of Snapchat and Instagram.

[19] Gabor Szabo and Bernardo A Huberman. 2010. Predicting the popularity of online content. *Commun. ACM* 53, 8 (2010), 80–88.