Predicting The Emotional Impact of a Research Article on Facebook

ABSTRACT

Social media platforms such as Twitter, Facebook, Reddit, and Google+ have a major influence in propagating information throughout society and make it easier for individuals to connect with each other. Whether it is sharing a post, posting a comment, uploading pictures, or reacting to a post; these interactions help people communicate with each other in a wide variety of ways. Sharing is a highly used feature that Facebook has and it facilitates the dispersion of many articles throughout the world. The emotional reactions an individual has to an article or post is a valuable data metric for researchers to analyze. The goal of this project is to predict the emotional impact of a research article that is shared on Facebook. This will be measured by collecting data on the reaction(s) of users, such as “Like, Love, Haha, Wow, Sad, and Angry” and the attempt to determine if there is a relation with other features of the article the post refers to. In general, organizations of various types as well as individuals interested in knowing the reaction(s) readers may have to a paper or article stand to benefit from this research. More importantly, authors of a publication may utilize this predictive model before posting their research on Facebook in order to ensure their research paper will not quickly develop a negative viral sentiment on social media. Other future work may include determining whether the author of the post has an influence on the elicited reactions. More importantly, future work related to this research would aim to further analyze text and determine if it improves the predictive model. The goal of this would be to provide authors with a sentiment score, which may help them predict the reaction of the community before formally submitting the research paper for peer review or final approval. Similarly, the institutions where the research was conducted may benefit; since it may help avoid negative publicity from a potentially offensive research paper.

Author Keywords

Altmetric; sentiment; emotional reaction; facebook reactions; facebook mining; social reaction; subject reaction

INTRODUCTION

Every researcher searches for credible and useful resources in order to further their own research, but finding those articles can be very time consuming especially when the author doesn’t know the impact of those articles beforehand. The use of social media platforms have skyrocketed and are taking over the internet. Articles are being shared every day on social media platforms and in order for researchers to utilize them; they have to analyze every single one of those articles to determine its benefit to their research. By seeing the reactions a similar research paper received to that of the one the author will publish may help them understand if they should make any changes before publishing. In addition, an author may be interested in knowing how the public will react to their research paper. Furthermore, it important to remember such a model may be detrimental to the research community in that may lead to self-censorship; where an author decides not to publish their research paper. This issue is further discussed in the discussion section. The goal of our project is to build predictive models that will predict the volume of each emotional reaction using a 3 class classification system represented by low, medium, and high. Two types of models will be used including both; Classification and Regression. The classification models used will be Random Forest, Naive-Bayes, and Decision Tree. For regression a multiple-linear regression model will be generated. Identify which features or combination of features are effective at building an accurate model to predict the emotional reactions of the audience; such as “love, haha, wow, sad, and angry.” The overall project is partitioned into multiple tasks, with the tasks split up as data pre-processing and cleaning followed by building our models.

RELATED WORK

A large amount of previous research has primarily focused on the predicting behaviors on Twitter [18, 19, 20]. Research conducted by Jérémie Clos and colleagues presented a technique to predict the emotional impact of news on consumers. They use a dataset extracted from the Facebook pages of the New York Times and mined positive or negative sentiment measured by likes or the emotional reaction [1]. Our research will go beyond New York Times articles and attempt to predict the reactions elicited from research papers. Another research paper used mined text as a feature and used a template based approach with 10 emotions (angry, worried, boring, depressing, happy, warm, odd, and informative.) Analyzing Table 3 in their research paper shows a list of words and phrases that elicited each particular reaction. The research that is presented here will differ in that it will use research publications as opposed to Chinese writings. Furthermore, these researchers did not use FaceBook reactions as their classifiers [4]. Some researchers have attempted to exploit the Facebook reaction feature to train a support vector machine classifier for emotion detection. They did this using a single classifier as well as feature combinations. The results of testing their model on existing benchmarks for emotion detection showed that there is room for improvement [13]. Another group of researchers collected posts (and their reactions) from Facebook pages of large supermarket chains and constructed a dataset. They improved the results of the neural networks by introducing a bootstrapping approach for sentiment and emotion mining on the comments for each post. Their final model (a combination of neural network and a baseline emotion miner) is able to predict the reaction distribution on Facebook posts with a mean squared error (or misclassification rate) of 0.135 [5]. Similarly, another group of researchers aimed to identify factors that affect reactions of customers to the posts that companies put up on their wall. Two local mobile game companies were analyzed and each post on each of the companies were coded in terms of structure, informativeness, type of reward, etc [9]. Although deep learning algorithms have been used to construct neural networks that are able to predict Facebook reactions to various types of posts. There is little to no research done on predicting the reaction to a post regarding a research paper using the Facebook reactions as classifiers. Other research has primarily focused on using the number of likes as a feature, this is most likely due to the reaction feature being released recently in February 24th of 2016. The earlier works done in regards to predicting emotions and likes provides our project with a great starting point, but we aim to further differentiate the inputs used. Our project will differentiate from the previous work by specifically predicting the emotional reaction of research publications. We will construct and utilize various features to generate our model. Initially, we will use the features that are easiest to extract and mine, and then further fine tune our model by using other features, potentially mined text in the form of bigrams.

DATASET

The dataset used for this project will be the altmetrics dataset provided to us for research purposes at Northern Illinois University. This dataset consists of articles and citations, which consists of features including; altmetric scores, #readers of various outlets, blogs, demographics, and where it has been cited. The entire altmetric dataset contained **8,165,874** records in total.

Dataset extraction

We utilized a JAVA application to extract necessary fields needed and used them as an input into a MySQL database. We extracted many features as we didn’t know which exact features we were going to use. While extracting the data and getting it processed we were making a few mistakes which we will get into. At first, we didn’t quite realize which features would be most important so we started off with only a select few. Once we figured out that we needed features that we didn’t initially include, we went back and processed all the data again. While looking at the dataset in excel after it was processed, we noticed that we only had 5 of the 6 reactions, missing the ‘love’ reaction data. We weren’t quite sure what happened so we went back to the code to fix whatever issue we were having and came to the conclusion that part of the code was commented out, resulting in only 5 reactions being processed which were ‘like’, ‘haha’, ‘wow’, ‘sad’, and

‘anger’. We further proceeded to uncomment it out and reprocess the data once again. Running into these errors and small mistakes took about 4 - 5 days to get everything right and situated. We thought it would be better to keep most of the features that we thought would be relevant and then start dropping them one by one as we created the model. We extracted a total of 29 features from the altmetric dataset, 6 emotional reaction features (“like, “love”, ”haha”, “wow”, “sad”, and, “anger”), a sum of all emotions, and then 5 normalized features ranging 0-1 for the 5 emotional reactions excluding “Like”. A problem we faced while creating our model was using the dataset. Once we normalized the data 0-1, some had 0 reactions but in the dataset it would be blank and not 0. While creating the model we got many errors saying that the model would not take any ‘NaN’ values. We were wondering why it would say that and when using python statements trying to replace those values ‘NaN’ to ‘0’, it would not work. So we had to go to the dataset and manually set those to 0 using Microsoft Excel. Doing that, solved our problem and we were able to continue with our model. As a starting point we looked closer into the Facebook attribute since this is where the basis of our proposal begins. This feature indicates how many times the article has been shared on Facebook. It also gives us the links to where it has been shared, which will be used in the Facebook Graph API. In addition, the data for the features indicated in the research hypothesis will originate from this dataset with the addition of Facebook reactions, which will be mined and concatenated to this existing dataset. One tool in particular that we are going to use is the Facebook Graph API to retrieve the reactions to the articles that have been shared on Facebook.

Dataset analysis

A quick analysis of the dataset found that the total counts of each of the reactions; love, haha, wow, sad, and anger were
134,357; 35,475; 87,976; 34,710; and 5,839,191 respectively. According to this observation we noticed that Facebook users are more likely to take the step to leave a reaction if it angers them. This makes sense in general, for example individuals are more likely to review a service if it was bad and less likely to review it if it was satisfactory or good. In addition, the total count for anger may be skewed by a few records in the dataset that may be considered to be outliers. Further analysis and processing limited the dataset to the top 20 subjects of which; medicine, science, nutritional sciences, social science, and life science represented the top five;

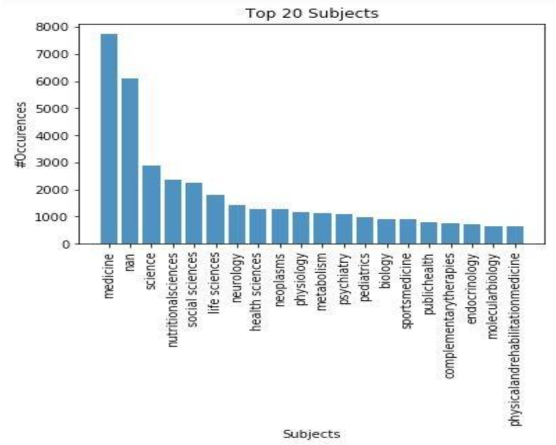


Figure 1: Showing frequency of each of the top 20 subjects in the dataset. *A larger* images is located in appendix A; figure 1.

Additionally, the counts of each emotional reaction by subject was charted, which may be found as Figure 2 in Appendix A. It shows that endocrinology, health sciences, and life sciences overall had a disproportionately higher count of anger reactions compared to the other reactions. This is likely because the elicitation of anger may be more influential in an individual voicing their opinion as a Facebook reaction. More detailed figures are shown in Appendix A, Figures 3 - 6. These show that the subject life science elicited the greatest count of anger reactions followed by endocrinology. For the reaction haha; medicine, science, and biology elicited the greatest counts. For the reaction love; medicine, science, and nutritional sciences. For sad; medicine, science, and life sciences. For wow; science, medicine, and life science.

API

In order to access the Facebook Graph API, it requires many steps need to be taken and we need user keys and application keys which can be found and created from the Facebook Developer website. In order to create and generate the access keys, one would need to create a developer account which then needs to be linked to a Facebook account. After that process has been done, an application has to be created on the developer account. You can only access Facebook information if the permissions have been set. To do that, access the Facebook Web Graph API and select every permission that you will need. Save that and it will create an access key which then you will be able to use in your python program. The access keys are unique to each and every person and should not be shared with anyone as they are very sensitive. The key contains access to your profile and anyone with the key can essentially do whatever they would like with your account. The access keys needed in order to retrieve data had to be generated every 2 hours because after it hits the 2 hour mark, the access keys would expire. Facebook has done this for security reasons. To regenerate the access keys, we needed to go back to the Facebook developer page. Facebook reactions became available to the public to use on February 24, 2016 and we took that in to factor and removed all the records from the dataset that were posted on Facebook before that date because it would not help us since our experiment was based on the basis of reactions. Retrieving the reactions from a Facebook post was not as simple as we originally thought. Facebook only had official documentation on how to use the Facebook Graph API for iOS, Android, Javascript, PHP, and Third-Party SDKs. We were not able to find documentation for python use as they did not officially support it. Our research mainly began on stackoverflow because many people were having problems on how to successfully connect to the Facebook API because Facebook did not officially support it. We went through different packages such as facepy, and facebook but figuring out the syntax was quite difficult because of the lack of documentation and information. Figuring out the syntax was a bit difficult while using those packages but we ran into a Facebook scraper that grabs all the reactions of every single post on a select page. We tried tweaking it so that it would grab only the post we wanted but it was quite difficult to do that because of the way the program was originally written. The output of the scraper would have the reactions of the post we wanted and posts we did not want. Also, the scraper took a very long time to process the data and time was not on our side. We eventually ditched this method once we got the Facebook package to to work with the right syntax. We achieved grabbing the reactions of a said post using the Facebook package with the /reaction endpoint and using version 2.7. There are many different versions out there but the /reaction endpoint came out in 2.6. We were originally going to use the version 2.6 but we thought it would be safer to use 2.7 because it was the newest build and many errors that were shown in previous versions have been fixed . In order to grab the reaction of a particular post, the object id had to be in a particular order, for example “PAGEID_POSTID”. Between the two different ways to grab the reactions of a said post, the scraper took over 10 seconds to process while using the stand alone Facebook package took 2 to 5 seconds per post. Processing the data using the Facebook package instead of the scraper cut our processing time by over half. The scraper took a very long time because it was grabbing reactions of every single post on the Facebook page while the stand alone grabbed reactions of a single post that we wanted. Using the Facebook package was definitely a time saver and a smarter route to go because it was grabbing the exact data that we wanted.

METHODS

In order to test our hypothesis that the subject would have the highest importance in determining the overall sentiment; two types of predictive models were chosen; classification and regression. The classification was done using three different classifiers including Random Forest, Naive Bayes, and Decision Tree. Classification included models utilizing the following algorithms; Random Forest, Naive Bayes, and Decision Tree. The 3 class classification cutoffs used to identify low, average, and high were <0.005, 0.005 - 0.4, and >0.4 respectively. Our multiple linear regression involved multiple independent variables which were continuous.

Feature Engineering

The models built using classification used the following combination sets of features; [Subject], [Subject; Facebook Count], [Subject; Twitter Count; mendeley_count; reddit_count], and [Subject; Twitter Count; mendeley_count; facebook_count; reddit_count]. The reason we did this was to see which set of features would give us the best accuracy and results at the end.

Regression

In statistics, there are two main types of regression. There is the simple linear regression and multiple linear regression. The linear regression is a linear approach for modelling the relationships between a scalar dependent variable ‘y’ and one or multiple explanatory variables denoted as ‘X’. One explanatory variable is called linear regression. As for multiple variables, this process is called multiple linear regression. In linear regression, the relationships are modelled using predictor functions where the model parameters are unknown and are estimated from the data. Most commonly, the conditional mean of ‘y’ given the value of ’X’ is assumed to be an affine function of ‘X’. And the opposite, the median or some other quantile of the conditional distribution of ‘y’ given ‘X’ is expressed as a linear function of ‘X’. Like all forms of regression analysis such as linear and multiple, linear regression is more focused on the conditional probability of ‘y’ and ‘X’, which is the domain of multivariate analysis. Multiple linear regression stems from linear regression and it the most common form of linear regression analysis. The multiple linear regression model is used to explain the relationship between one continuous dependant variable and two or more independent variables. The independent variables can be continuous or categorical. In this experiment we conducted on prediction the emotional impact of an article shared, we used the multiple linear regression model and categorical models as well. Before pursuing the multiple linear regression, we did the Random Forest model before all other models. We thought it would be a great way to get a spectrum of all models to see which one works out best. While performing the multiple linear regression model, we dropped all the columns that were not needed such as the journal name, ‘titleofpaper’, ‘publishdate’, ‘latestpost’, ‘facebook_id’, ‘doi_id’, ‘alt_id’, and all the reactions except the one we were trying to predict. For the subject, we used a label encoder to give each subject a unique number. Splitting the data, we used the same random state ‘0’, and the test size of 20% of the entire dataset we were currently using. While predicting, we believe we were having an issue of overfitting. Overfitting is a concept and is a modelling error which occurs when a function is too closely fit to a limited set of data points. The reasoning behind this is because we were getting a mean absolute error of ‘0’, the mean squared error of ‘0’, and R squared of ‘1’ as well for each reaction that was being predicted.

RESULTS

The results varied between each of the different types of models used as well as the models of each type. In regards to classification; overall, the Random Forest and Naive Bayes model outperformed the Decision Tree. In general the predictive accuracy of the love reaction underperformed that of others. The feature importances in general varied by the feature set combinations analyzed, but in general the features representing counts of the various social media metrics outperformed subject.

Feature importances

One of the results of building the Random Forest model provided data on the feature importances for each of the features used in the four feature sets that were tested. For the features combination including subject and Facebook count only; the feature importances were typically between .75 to .80 and 0.2 to 0.25 respectively. This set of features was used as a way to validate our model for correctness, since it was assumed that the greater number of Facebook posts made referencing a research paper would have an impact on the reactions. This was proved true as the importance of Facebook count significantly outperformed subject. For this set of features the feature importances were approximately similar for all the reactions.

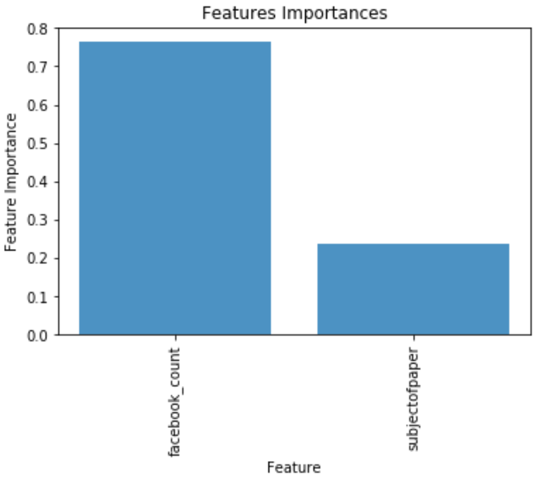


Figure 7, showing feature importances of each feature contained in feature set 1(Similar for all reactions) *A larger* images is located in appendix A; figure 7.

For the feature combination including mendeley count, twitter count, reddit count, and subject the feature importances were approximately the same for all the reactions at approximately 0.30 to 0.35, 0.37 to 0.40, 0.08 to 0.1, and 0.2 to 0.23 respectively.

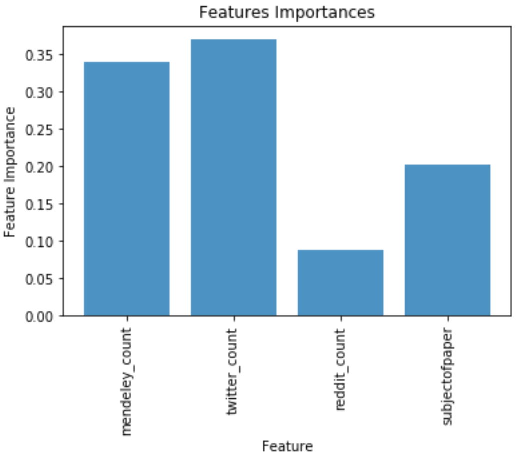


Figure 8, showing feature importances of each feature contained in feature set 2 (Distribution similar for all reactions) *A larger* images is located in appendix A; figure 8.

For the feature set adding Facebook count to the previous set resulted in the feature importance varying slightly for certain groups of reactions. The similarity of feature importances were fairly similar for love, haha, and wow at approximately 0.2 to 0.23, 0.18 to 0.2, 0.4 to 0.44, 0.5 to 0.7, and 0.1 to 0.15 for mendeley count, twitter count, FaceBook count, reddit count, and subject respectively.

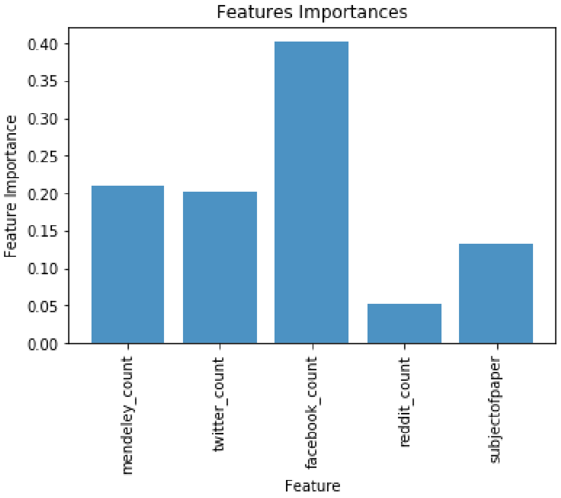


Figure 9, showing feature importances of each feature contained in feature set 3 for reactions (Distribution similar for love, haha, and wow). *A larger* images is located in appendix A; figure 9.

As opposed to the other feature sets analyzed this set varied for the reactions of sad and anger, which had importances of approximately; 0.2, 0.23 to 0.25, 0.38 to 0.4, 0.5 to 0.7, 0.1 to 0.15 respectively. The slight difference that occurs here is that adding the Facebook count as a feature increases the feature importance of twitter count and subject of paper while slightly reducing that of reddit count and subject as may be observed by comparing figure x and x.

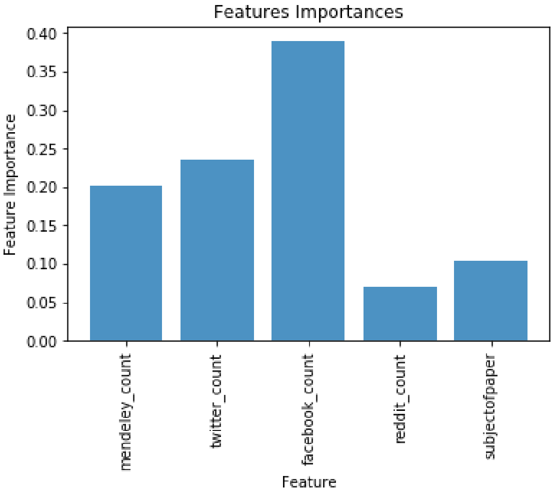


Figure 10, showing feature importances of each feature contained in feature set 3 for reactions (Distribution similar for sad and anger). *A larger* images is located in appendix A; figure 10.

Feature accuracies

Another important metric to analyze is the accuracy of the model in predicting each of the reactions. Overall, the distribution was similar for each of the feature sets explored. In general, love had the lowest accuracy at approximately 76% to 78% followed by the reaction wow at approximately 82% to 85%. The accuracy of predicting the reactions sad, anger, and haha were similar and greater than 94%.

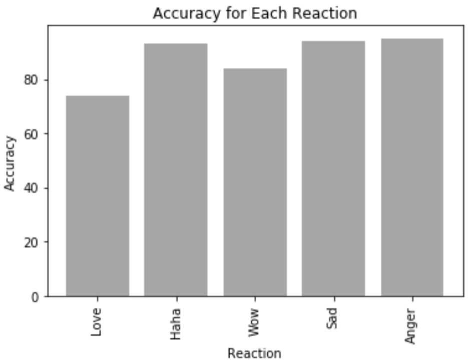


Figure 11, showing predictive accuracy of each reaction using Random Forest Model. *A larger* images is located in appendix A; figure 11.

Regression

The regression model proved to be very accurate, but the accuracy appeared to be too perfect as further analyzing the actual vs predicted values always showed a difference of one and typically in the same direction.

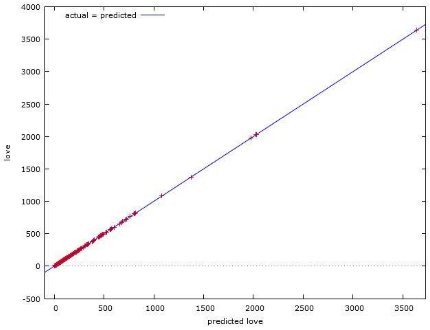


Figure 12, showing actual vs. predicted of the multi-linear model (Love reaction shown). *A larger* images is located in appendix A; figure 12.

Overall, the results showed that for classification models predicting love had the lowest accuracy, with anger, sad, and haha with the highest with wow falling between the aforementioned groups. The Random Forest and Naive Bayes consistently outperformed Decision Tree. The regression model outperformed all the classification models, although further analyzing the results leads to suspicion of overfitting and it is uncertain whether the model would would be usable in predicting reactions in a real world use case. Further testing and validation would be required to conclusively determine if it was superior than classification.

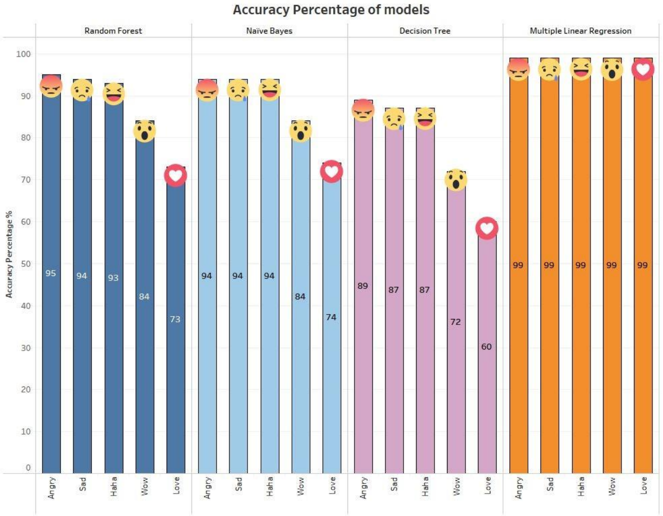
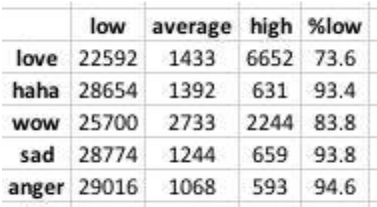


Figure 13, showing a summary of the accuracy for each reaction of each model; left to right (Random Forest, Naive Bayes, Decision Tree, Multiple Linear Regression). *A larger* images is located in appendix A; figure 13.

DISCUSSION

It is important to note that there may have been several biases introduced into the project during several stages from building the initial dataset to generating the models. It is noteworthy to mention that a significant number of records had more than one subject and it was assumed that the first subject in the original dataset was the primary subject, this was an assumption that was made. Furthermore, there were two outliers in the dataset. One of which was a record represented by the subject 'life sciences' and the other by 'endocrinology'. Both these records had a disproportionately higher count for the anger reaction compared to the rest of the records in the dataset. An aspect to consider regarding these outliers is overall human behavior. Typically, humans are more likely to react if they are angered or saddened by something. For example, users are more likely to leave negative reviews for a product or service if they are dissatisfied than a positive review if they are satisfied. This is because dissatisfaction causes undue stress and anxiety. This was observed in our data, where the total counts of anger greatly exceeded that of all the other reactions combined. The counts of all the other reactions combined was approximately 292,000 while that of anger alone was approximately 5,800,000. Therefore, the counts of anger alone represented 95% of the total counts of all reactions in the dataset. As mentioned earlier there were two significant outlier records that represented a disproportionately high count for the anger reaction at approximately 3,900,0000 followed by another record representing a total count of the anger reaction at approximately 1,000,000, while having 0 for all the other reactions. The subjects of these records were life science and endocrinology respectively. It is interesting to note how all the reactions other than anger were 0. If the counts are adjusted to remove these records, then 75% of the total reaction counts are represented by anger. It is important to note that even when correcting the outliers, anger still represents a disproportionately higher count than all the other reactions combined. This again is likely the result of human nature where anger elicits greater reactful actions. Although this is an important insight into the dataset, these outliers should not have had an influence on the accuracy of our models, since all reactions were normalized to 0 - 1, represented as the proportion of the particular reaction to the total count of reactions for that particular record. The results presented in this paper represent using a 3 class classification. To reiterate, the cutoffs used for low, average, and high were <0.005, 0.005 - 0.4, and >0.4 respectively. The reason the ranges were skewed closer to 0 is because the counts of reactions in the low class disproportionately exceeded that of the other two classes. For example, the classes for the sad reaction were split as 28,774; 1,244; and 659 for each of the classes low, average and high respectively. As is obvious even using the highly skewed cutoff ranges the low class represented a higher count of reactions. This pattern followed for all the other reaction. Table x shows counts for the reactions.



	low	average	high	%low
love	22592	1433	6652	73.6
haha	28654	1392	631	93.4
wow	25700	2733	2244	83.8
sad	28774	1244	659	93.8
anger	29016	1068	593	94.6

Table 1 frequency counts of classes for each reaction.

It is important to note that even when using a more evenly split cutoff, it did not affect the accuracy of our model(s). It is unknown if it would have an effect on other models though and believe it is noteworthy to mention this statistic for any researcher continuing this work. The results of the classification models seemed more realistic compared to those generated by the regression model. It is important to note that analyzing the predicted vs actual values generated by the model shows that if the predicted did not match the actual exactly it was off by -1. The odd thing about this is that it was never off by +1. This may have been due to chance, although these results were consistent each time the model was run. In general the predicted values were normally exactly the actual, again this was suspect to being too perfect and it was determined further modifications would need to be made. It was concluded that the regression model was overfitting and therefore giving us an accuracy of 99%. It is noteworthy to remember that out of the 42% of records processed from the entire dataset, 98% were removed since they contained no Facebook posts or posts made prior to February 2016. Further, and additional 4 - 5% were removed because the post or author page had been deleted. Although the dataset used for building the models contained approximately 30,678 records, it would be beneficial to process the remaining 58% of the raw altmetric dataset. This would likely add and slightly over 40,000 records giving us a dataset of approximately 70,000 to use as input in generating our models. It is unknown whether this would improve the accuracy of the model, but given the fact that Facebook reactions have only been available since February 2016, it would likely provide a more diverse dataset. In addition, as time progresses more data will be reaction data will be available. This data will become more diverse as time progresses, since it may include research from subjects that are still are in the process of producing results. Many of these subjects are likely going to have an emotional impact on individuals, since they will be related to artificial intelligence, synthetic life, modifying human DNA, and other subjects deemed to be playing god by the general population. Furthermore, using a larger dataset may provide us the ability to generate models that are accurate using other features such as the title of the research paper or article or the text from the abstract. In order to build an accurate model from such sparse and varied text data would be most beneficial using the largest dataset possible.

CONCLUSION

In this experiment, we presented an approach for predicting the emotional impact that an article can have on an audience using different models such as Random Forest, Decision Tree, Naive Bayes Classifier, and Multi-Linear Regression. Both Random Forest and Naive Bayes Classifier performed well while Naive Bayes Classifier performed slightly better than Random Forest with the Decision Tree underperforming both. We focused on the Facebook reactions; "Like, Love, Haha, Wow, Sad, and Angry" but only predicting all the reactions except 'like' and classifying them as low, medium, high. There are many different methods to develop a prediction model, but based on our dataset we believe that this was the best way to approach this problem. We did not use the bootstrapping approach but we believe if we did, it would have helped reduce the overfitting issue we were having with the Multi-Linear Regression. Facebook came out with reactions in 2016 and that is why we believe that not much work has been done in the past on Facebook reaction predictions due to lack of reactions to a post. We hope that this approach will help authors and publishers see the volume of reactions of an article before actually posting and sharing it on Facebook. Facebook reaction predictions can give a better understanding on how an article will do without the risk. Since social media has taken everything and everybody by storm, Facebook reaction prediction is the best way to figure out the future for an article that has been posted. It would be interesting to identify if certain words or phrases found in the abstract or certain sections of the paper were able to be able to predict the overall reactions of a research paper. It is likely that using the subject, title, and text of the abstract could be the best features in order to make this determination. Although, our models showed that the subject feature was not as important as the counts of posts from the various social media outlets. Intuitively it seems as though the text should be able to be used in the predictive model using NLP neural networks. The reason for this assumption is that humans read the text and then the reaction is elicited and since there has to be a pattern to this it has to be possible to to this through machine learning and analyzing existing data. This would be the holy grail for scientists and researchers since it would help them predict how the general scientific community and public as well would react to the research paper. Knowing this information may help them decide if they want to "market" their research to a great extent. Unfortunately, a downside of a predictive model that allows researchers to ascertain the emotional sentiment is self-censorship by the authors and researchers of a paper or article. This is because if the model predicts the paper will receive a negative sentiment represented by anger or sad reactions, the authors may choose not to publish their research. This may have a negative consequence in making discoveries as sometimes the best and most innovative research is not received positively by the general public. As evidenced in our dataset it was obvious there were a higher number of anger and sad reactions to genetically modified organisms and anyone in the scientific community knows that this is indeed the future we as a society are moving in. With the human genome mapped out over a decade ago, computing hardware finally exists now to allow researchers in fields such as bioinformatics to analyze that data at a greater rate. High throughput DNA sequencing technologies is also increasing the amount of genetic data from other organisms such as plants and non-humans that may be analyzed against the human genome. This research is without doubt going to advance society and mankind as a whole, but in the process will receive a lot of pushback from conservative individuals and those who lack basic scientific knowledge of computer science and biology.

FUTURE WORK

As for future work, we believe that this would be a great stepping stone for a researcher that would like to perform this in a different approach and in all hopes that this will help further their own work. There is a lot more work needed to be done to get a better accuracy in prediction. As time goes on, more and more people will start to use the reactions, meaning that there will be better data to work with since reactions came out in February 24, 2016. With the new dataset, retesting the models would be a great idea to see how it performs. In our experiment, we did not implement or use any text features so for future work, extracting text from the abstract of the article, title, Facebook comments and twitter comments can potentially help the outcome. The reason is because we believe mining the text would give us a better idea of the reactions of a post from user comments. The Facebook author as a feature would be a great feature as well because different users have different amount of friends list and popularity does have an effect on how much exposure an article gets. The secondary, tertiary subjects could have an effect with the accuracy if used. Another thought was to derive reactions from Love, Haha, Wow and to derive reactions from Sad and Anger. Love, Haha, and Wow generally all refer to one thing which is a positive emotion while Sad and Anger refer to a negative emotion. The last feature that we believe that can help with future work is the date feature if future researchers are trying to predict the exact number of reactions a certain post on Facebook gets. Depending on when it was published, it can give us insight on how long the article was posted for or shared, and the impact it had. There is a lot of work to be done with this idea and can be improved in many different ways. Additionally, being able to use the regression model would be of great importance in solving the problem at hand of being able to provide a quantitative prediction of the reactions a post may elicit and would certainly be of benefit to the community to further improve and fine tune the regression model to reduce the degree of overfitting that it exhibited. In addition, potentially testing out models beyond the ones explored in this paper may help generate more accurate models. Despite the well intentions of this research and proposed future work, the researchers furthering this model should remain cognizant that a perfect model may lead to self censorship by authors and researchers as mentioned in the discussion.

REFERENCES

[1] J. Clos, A. Bandhakavi, N. Wiratunga, and G. Cabanac, "Predicting Emotional Reaction in Social Networks," in *Advances in Information Retrieval*, 2017, pp. 527–533.

[2] S. Tettegah, *Emotions, Technology, and Social Media*. Elsevier Science & Technology Books, 2016.

[3] K. H.-Y. Lin, C. Yang, and H.-H. Chen, "What emotions do news articles trigger in their readers?," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 733–734.

[4] Y.-C. Chang, C.-H. Chu, C. C. Chen, and W.-L. Hsu, "Linguistic Template Extraction for Recognizing Reader-Emotion," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 21, no. 1, pp. 29–50, 2016.

[5] F. Krebs, B. Lubascher, T. Moers, P. Schaap, and G. Spanakis, "Social Emotion Mining Techniques for Facebook Posts Reaction Prediction," *arXiv [cs.AI]*, 08-Dec-2017.

[6] T. Vepsäläinen, H. Li, and R. Suomi, "Facebook likes and public opinion: Predicting the 2015 Finnish parliamentary elections," *Gov. Inf. Q.*, vol. 34, no. 3, pp. 524–532, Sep. 2017.

[7] M. Söderqvist, "Social Media Reactions-An Empirical Study about the Shifting Communication Dynamics on Facebook," 2016.

[8] C. T. Carr, R. A. Hayes, and E. M. Sumner, "Predicting a Threshold of Perceived Facebook Post Success via Likes and Reactions: A Test of Explanatory Mechanisms." Routledge, pp. 1–11, 2017.

[9] H. Jeon and H. J. Ahn, "Identification of the factors that affect the user reaction to posts on Facebook brand pages." Institute of Electrical and Electronics Engineers Inc., pp. 203–206, 2015.

[10] S. Ahmed and M. Haag, "Emoji-what is their purpose? A study of effective communication on facebook." 2017.

[11] B. V. Srinivasan, A. Natarajan, R. Sinha, V. Gupta, S. Revankar, and B. Ravindran, "Will your facebook post be

engaging?," in *Proceedings of the 1st workshop on User engagement optimization*, 2013, pp. 25–28.

[12] Y. Tian, T. Galery, G. Dulcinati, E. Molimpakis, and C. Sun, "Facebook sentiment: Reactions and Emojis," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 2017.

[13] C. Pool and M. Nissim, "Distant supervision for emotion detection using Facebook reactions," *arXiv [cs.CL]*, 09-Nov-2016.

[14] R. A. Calvo and S. Mac Kim, "EMOTIONS IN TEXT: DIMENSIONAL AND CATEGORICAL MODELS," *Comput. Intell.*, vol. 29, no. 3, pp. 527–543, Aug. 2013.

[15] G. W. Tigwell and D. R. Flatla, "Oh that's what you meant!: reducing emoji misunderstanding," in *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, 2016, pp. 859–866.

[16] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter." pp. 149–156, 2011.

[17] S. Madisetty and M. S. Desarkar, "An ensemble based method for predicting emotion intensity of tweets," vol. 10682 LNAI. Springer Verlag, pp. 359–370, 2017.

[18] T. Hasegawa, "Predicting and eliciting addressee's emotion in online dialogue," *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 964–972, 2013.

[19] J. Chen, Y. Liu, and M. Zou, "User emotion for modeling retweeting behaviors," *Neural Netw.*, vol. 96, pp. 11–21, Dec. 2017.

[20] S. M. Mohammad and S. Kiritchenko, "Using Hashtags to Capture Fine Emotion Categories from Tweets," *Comput. Intell.*, vol. 31, no. 2, pp. 301–326, May 2015.

Appendix A:

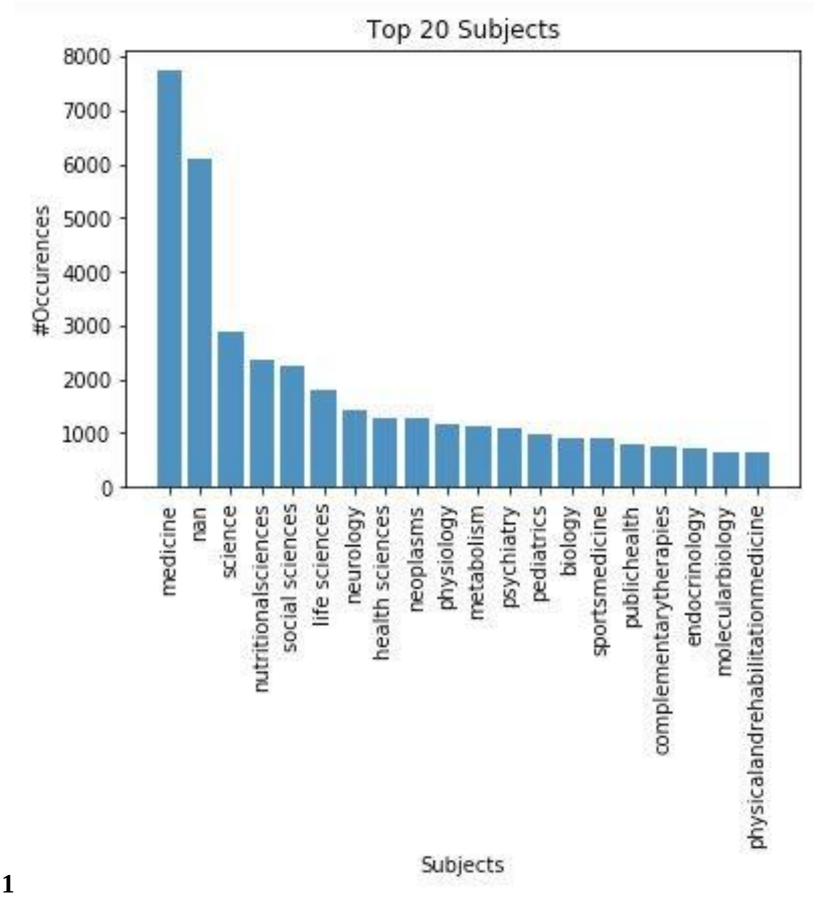


Figure 1

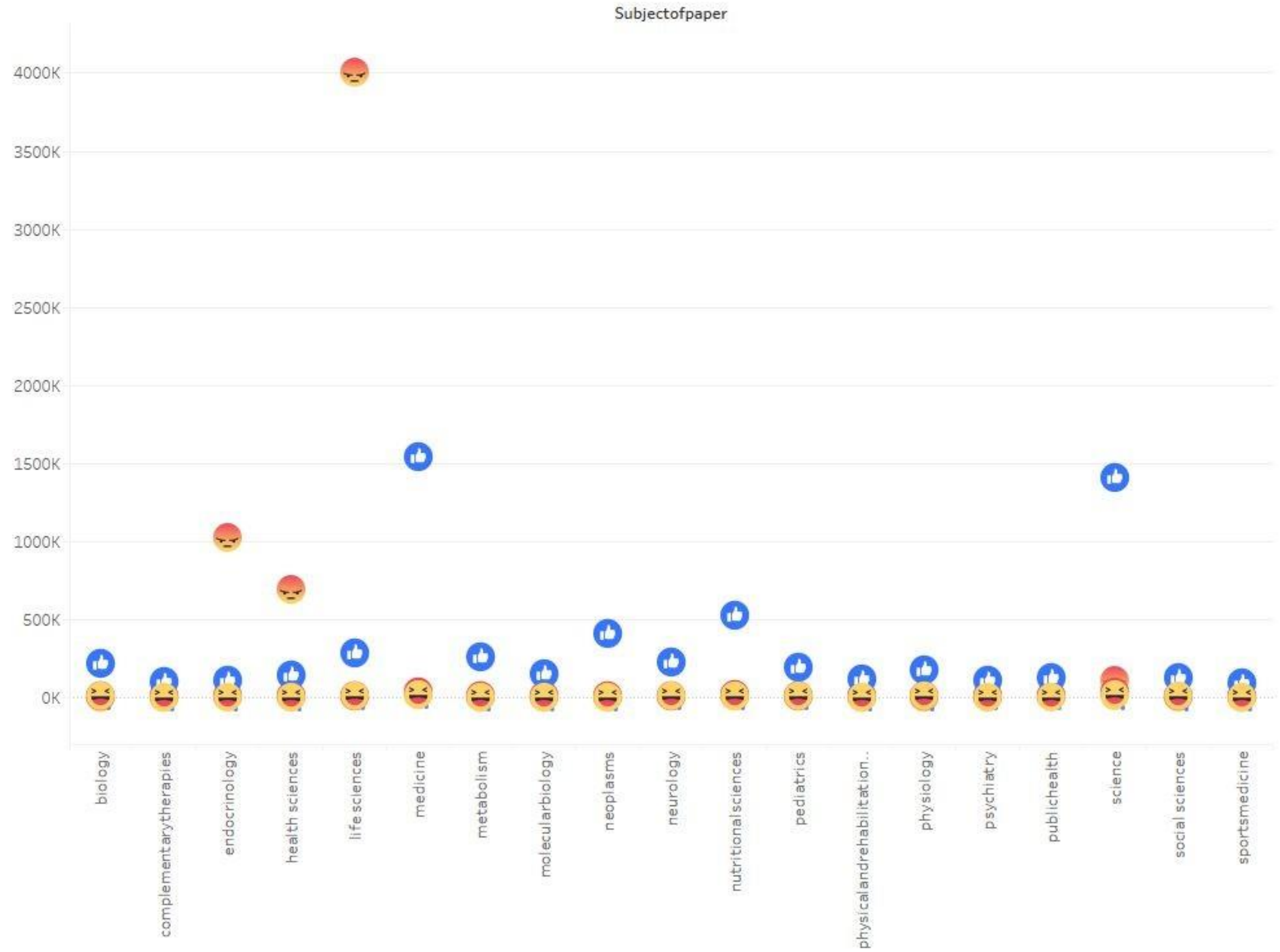

Subjectofpaper
4000K
3500K
3000K
2500K
2000K
1500K
1000K
500K
0K
biology
complementarytherapies
endocrinology
health sciences
life sciences
medicine
metabolism
molecularbiology
neoplasms
neurology
nutritionalsciences
pediatrics
physicalandrehabilitation..
physiology
psychiatry
publichealth
science
social sciences
sportsmedicine

Figure 2

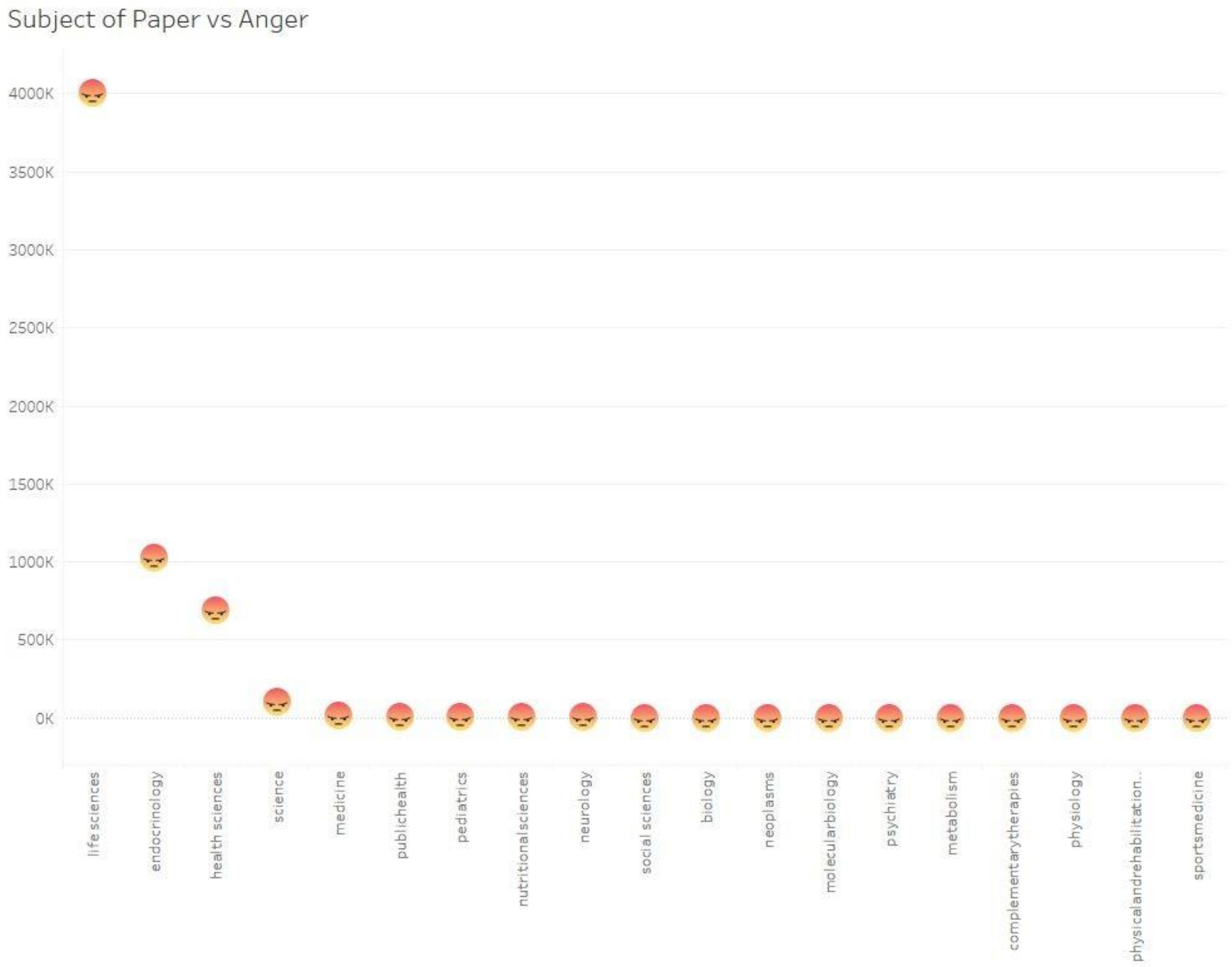

Subject of Paper vs Anger
4000K
3500K
3000K
2500K
2000K
1500K
1000K
500K
0K
life sciences
endocrinology
health sciences
science
medicine
publichealth
pediatrics
nutritionalsciences
neurology
social sciences
biology
neoplasms
molecularbiology
psychiatry
metabolism
complementarytherapies
physiology
physicalandrehabilitation..
sportsmedicine

Figure 3

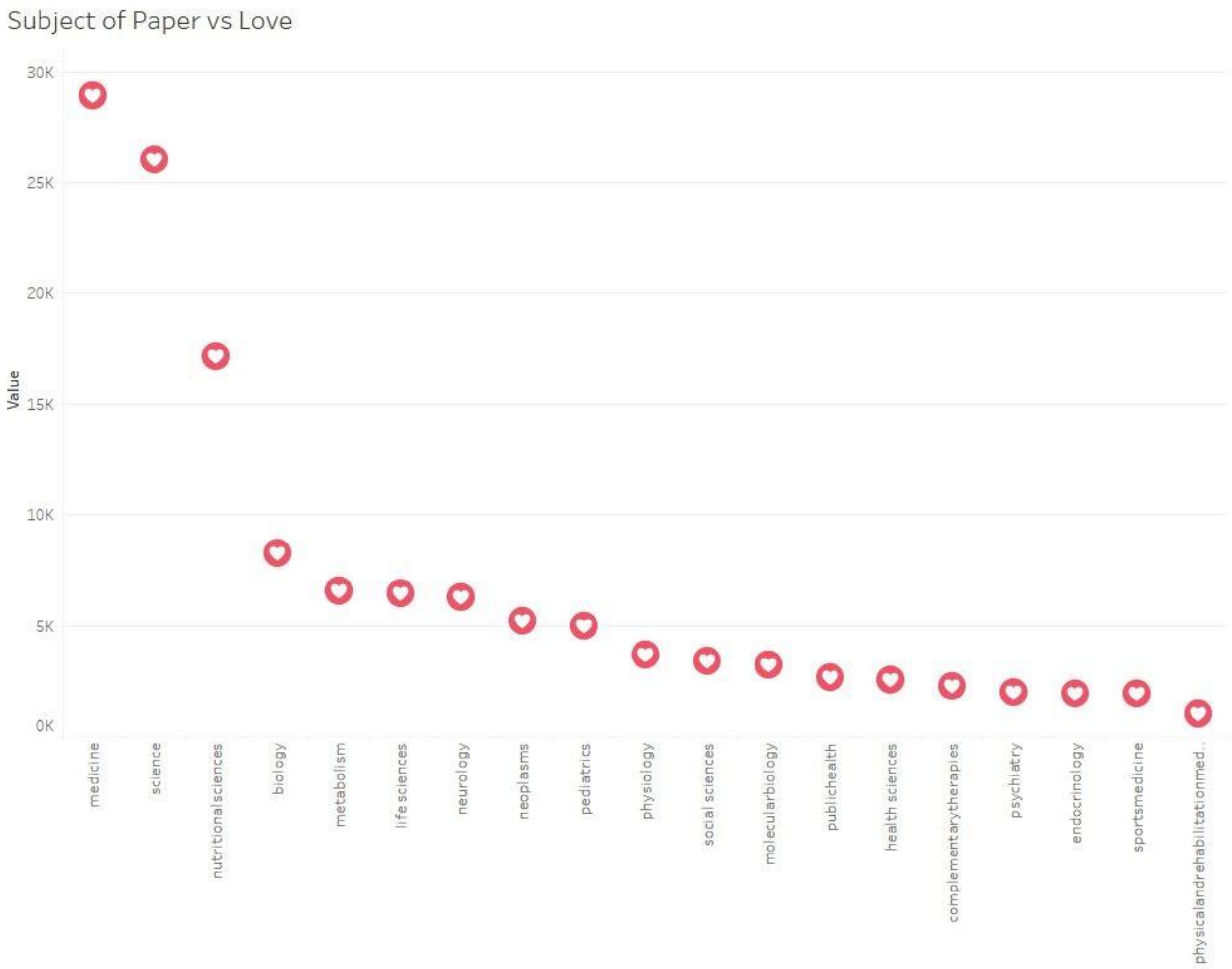

Subject of Paper vs Love
Value
30K
25K
20K
15K
10K
5K
0K
medicine
science
nutritionalsciences
biology
metabolism
life sciences
neurology
neoplasms
pediatrics
physiology
social sciences
molecularbiology
publichealth
health sciences
complementarytherapies
psychiatry
endocrinology
sportsmedicine
physicalandrehabilitationmed..

Figure 4

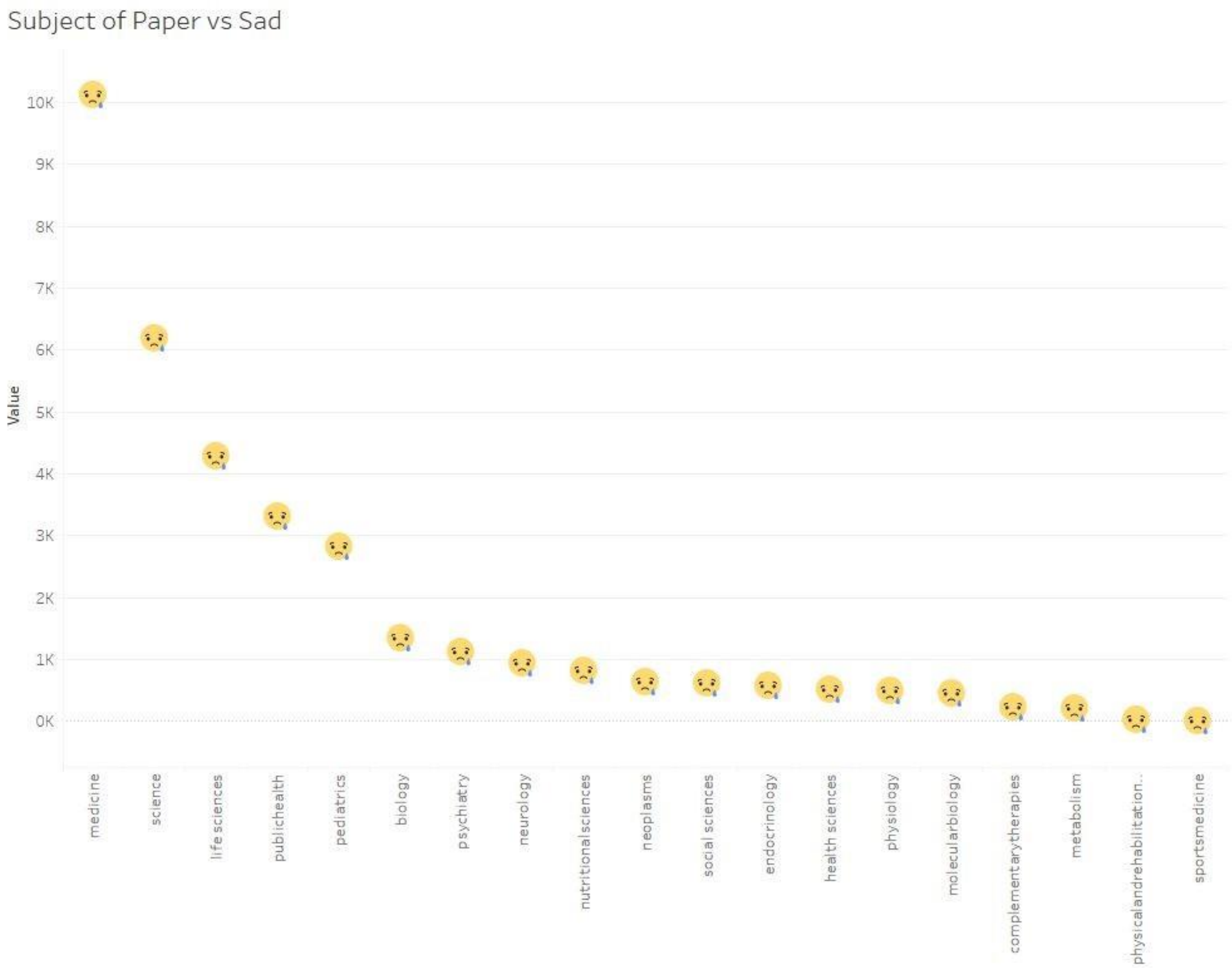

Subject of Paper vs Sad
Value
10K
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
medicine
science
life sciences
publichealth
pediatrics
biology
psychiatry
neurology
nutritional sciences
neoplasms
social sciences
endocrinology
health sciences
physiology
molecularbiology
complementarytherapies
metabolism
physicalandrehabilitation..
sportsmedicine

Figure 5

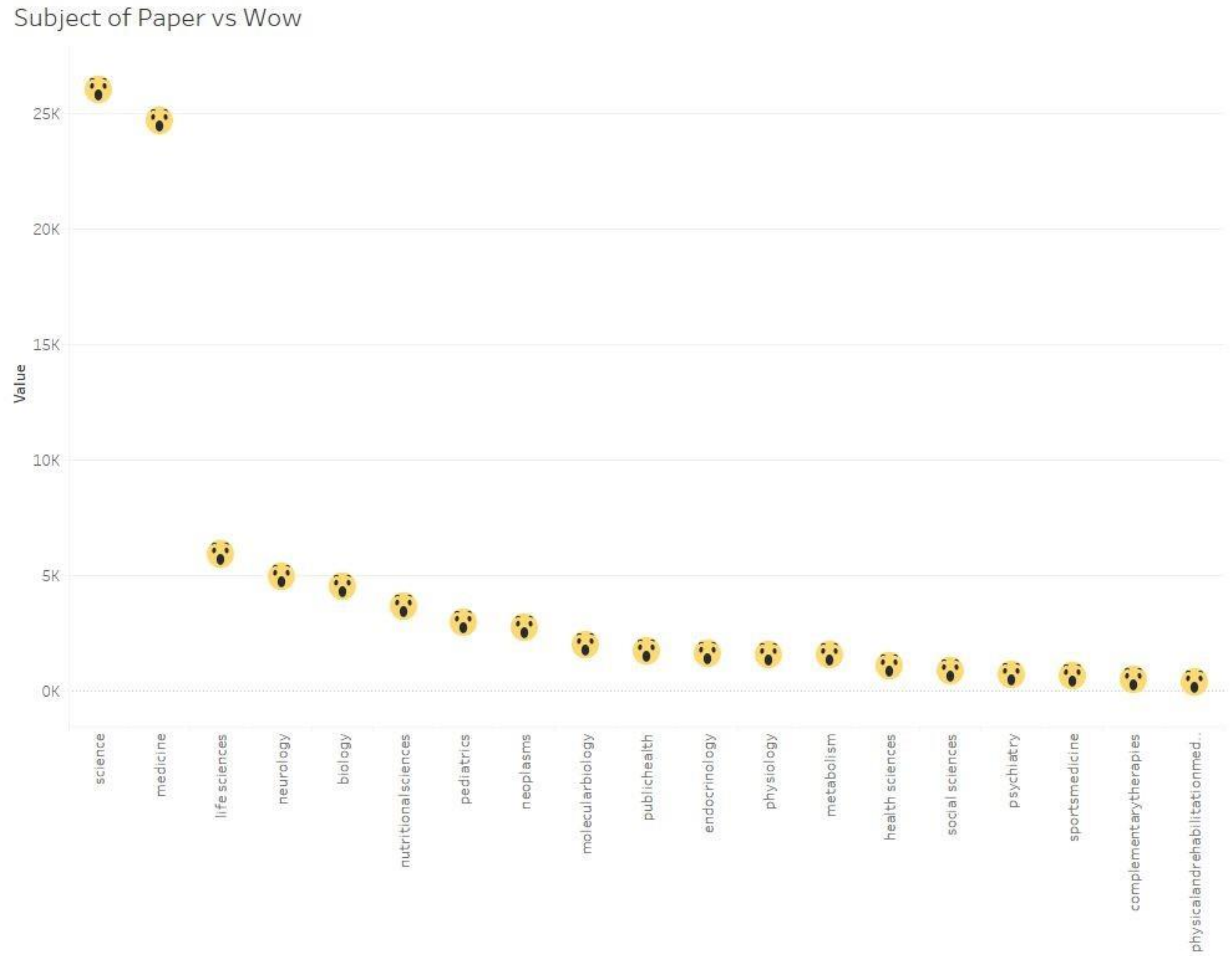


Figure 6

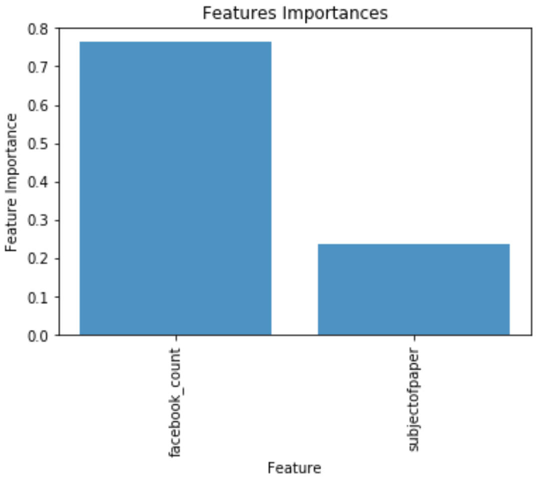


Figure 7

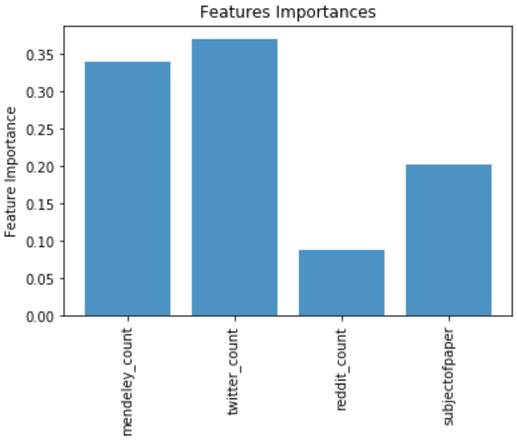


Figure 8

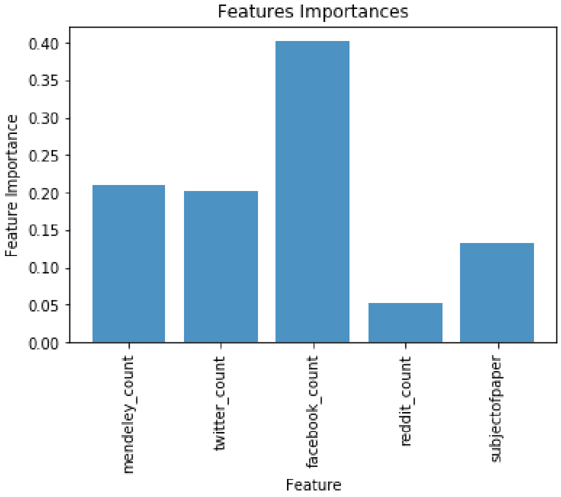


Figure 9

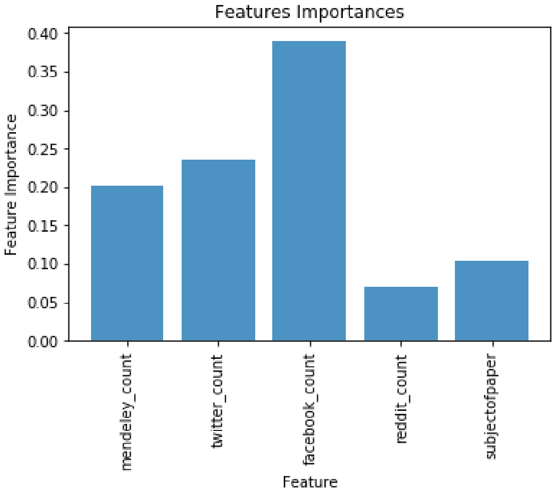


Figure 10

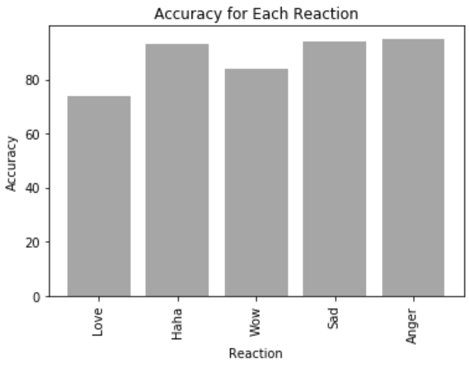


Figure 11

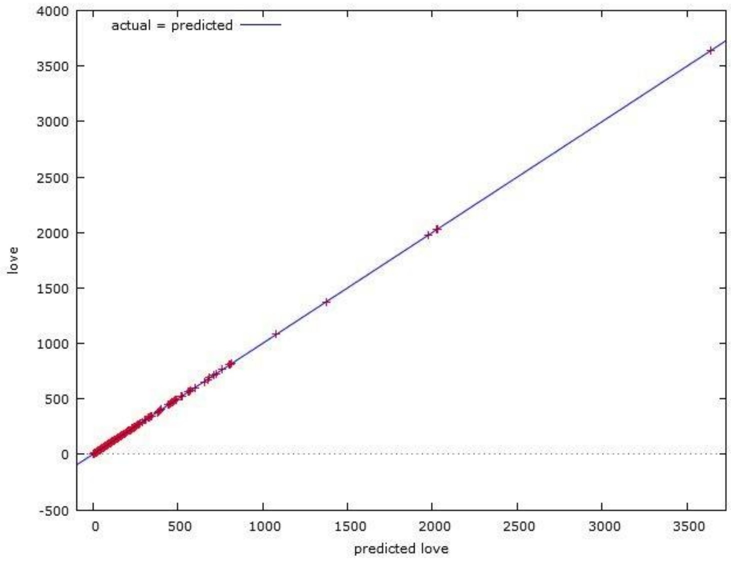


Figure 12

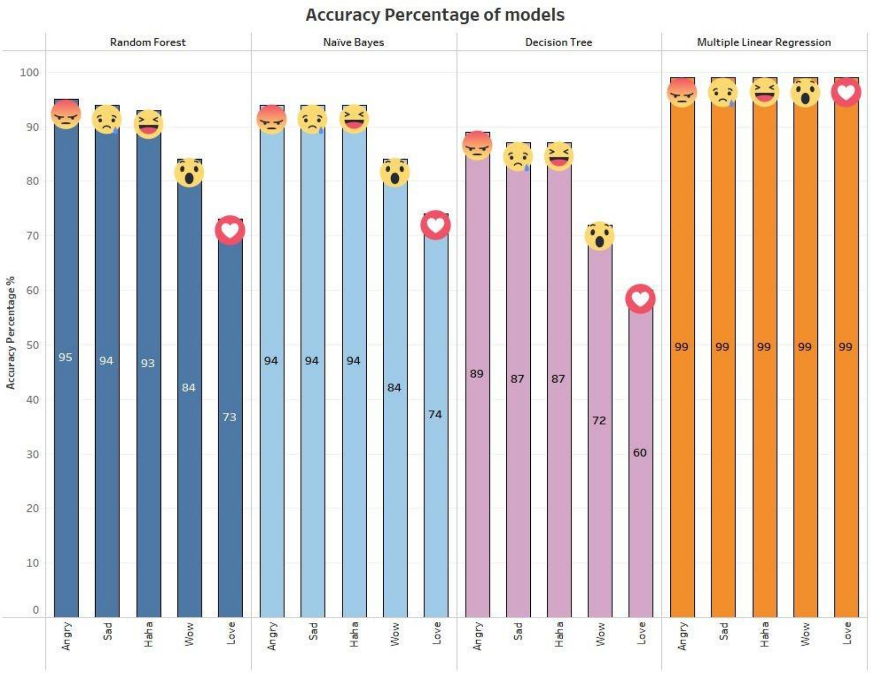


Figure 13