Finding a Related News Article of Posts in Social Media: The Need to Consider Emotion as a Feature

ByungSoo Ko, Chanyong Park, Dongkeon Lee, Jaewon Kim
School of Computing, Korea Advanced Institute of Science and Technology (KAIST)
291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea
{kobiso, ptparty, hagg30, jwkim1993}@kaist.ac.kr

***Abstract*—As social media data grows to tremendous size, understanding of posts in social media becomes critical for plenty of applications such as political or commercial analysis. It is helpful by giving us an insight into how the significant social issues affects one person or a group of people. Therefore, in this paper, we want to find a news article that relate to the users' post on the Facebook. We have observed that keyword only models do not work well for similar keyword, but we found that lots of posts have distinguishable difference in their emotional distribution. Therefore, we suggest the use of emotion features as a solution to this. By comparing the performance of classification model with emotion and no emotion that predict a related news article from the comment, we show that the classification model with emotion is acceptable technical approach for the given problem.**

Keywords*—*text classification, sentiment analysis, social media

I. Introduction

As smartphones has been used widely, many people can access various information and news through the internet. Furthermore, most of the news were delivered by website of each news company a few years ago. However, social media has taken charge of the role these days. This is because, it is easier to access and express their thought and emotion on the social media about each news article. In this circumstance, finding a news article that relates to the user's post in social media is important. There are two reasons why it is necessary. First, people can know what news article their friends are interested in and it will be helpful to communicate with their friends. Second, from the macro perspective, identifying the news articles that people are most interested in helps us understand the social trend. Therefore, we try to find a news article relate to the users' posts in social media. Example of it is shown in Figure 1. Each post on timeline will be classified to a related news article.

In order to achieve this, simple method that can be applied is to use similarity of keyword vectors of posts and news articles. However, if there are many news articles with similar keyword vector, it will be hard to classify into correct news article that invoked the post. As shown in Figure 2, assume that there are two news articles that seems to be related with the given post by keyword similarity. In fact, the news article associated with the post is news article 1. However, the cosine similarity between news article 2 and the post vector is 20.7%, which is higher than that of news article 1 and the post vector, 14.5%. Therefore, this example shows the keyword only method is not enough. Thus, we need more features to get better matching between the news articles and the posts.

Figure 3 shows people's reaction to different news articles. We collected reaction data from four news pages on social media. Figure 3a is related to United Airlines that dragged off a passenger, and Figure 3b is related to what pope said. As you can see, most people express "love" emotion on positive news article and "angry" on negative news article. So if the news articles influence emotion of users, emotion of each post in social media also can be used as a feature to find a related news article of the post. Also we use the date when the post was uploaded as a feature. According to MarketingProfs[1], more than 2 million articles are published everyday on the web and people tend to focus on recent news articles rather than old news articles. So it is natural to classify a post to the news article which is posted near the date when the post is uploaded.

For clarification, the term "post" collectively refers to as all texts written by users on social media. And the term "comment" refers to texts written by users under each news article. So, in the case of comment, we can easily map the comments and news articles. And this is the reason why we use comments for training data of our model, not posts.

In this paper, we present three classification models to find a related news article of posts in social media. We conduct experiments with deep neural networks (DNNs) by differentiating inputs among comment, emotion and date. To our knowledge, this is the first reseach classifying news articles from the comments by using emotional information to give more relevant and accurate classification result.

II. Related Work

In fact, there are few studies using personal emotional information to find news articles. However There are many research that can be helpful to our goal. There has been extensive work on emotion extraction and classification with emotion information that show using emotions for feature can help the performance of the model. We briefly describe some of the related work in this section.

A. Emotion Extraction

There have been many studies on emotion analysis which classify sentiment from the text. And there are two main ways to it, first one is lexicon based method and second one is machine learning based method. Lexicon-based method uses dictionary of words annotated with their semantic orientation

All four authors contributed equally to this work

[1]https://www.marketingprofs.com/

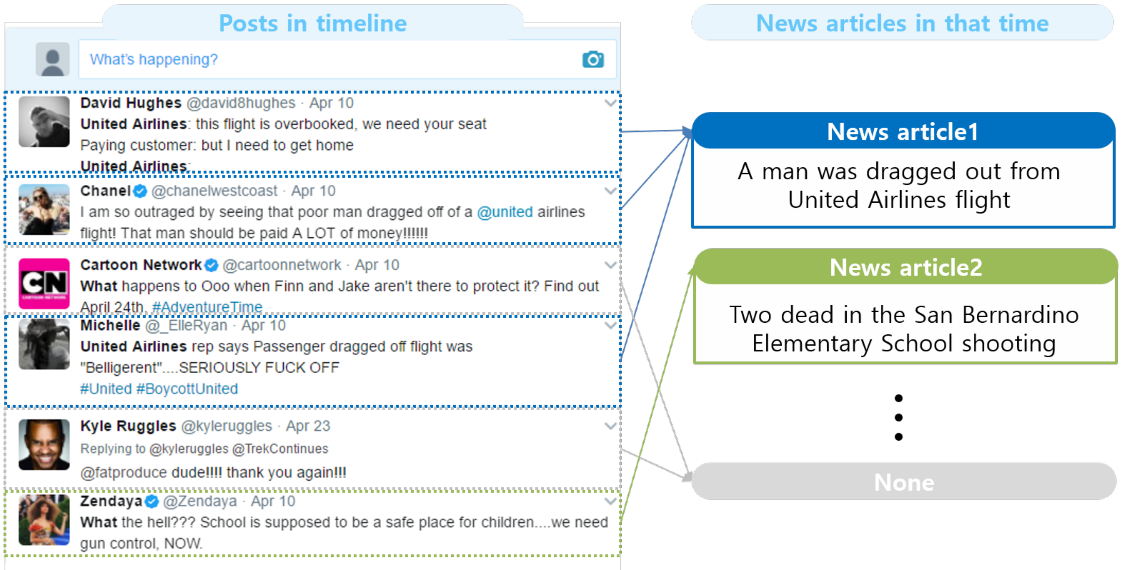
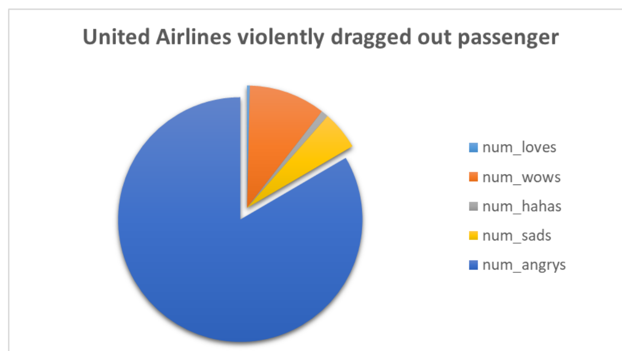
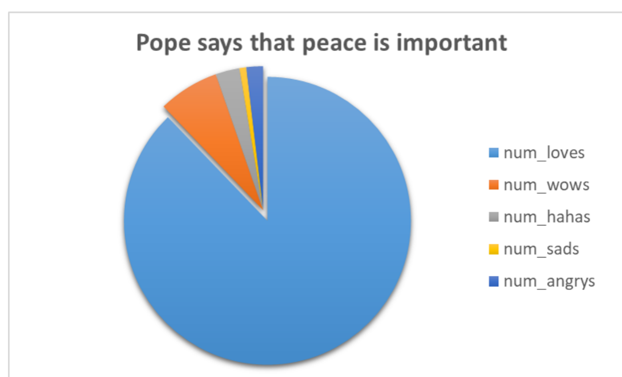


Fig. 1: Example of Extracting News Article from Posts in Timeline

Fig. 2: Example of Difficulty in Keyword Only Model



(a) Emotion Distribution of Negative News Article



(b) Emotion Distribution of Positive News Article

Fig. 3: Emotion Distribution of News Articles

and incorporates intensification and negation. M. Taboada et al. [1] is generally considered the principal work on using lexicon based method. By contrast, machine learning based method uses machine learning algorithms of text classification for sentiment analysis. And the principal work on machine learning based method is B. Pang et al. [2]. They build model using support vector machines (SVMs) and Naive Bayes (NB). However, in recent studies M. Ghiassi et al. [3] and S. Poria et al. [4] did, artificial neural networks are commonly used for the method and outperform the SVM and NB approach.

B. Classification with Emotion Information

J. Herzig et al.[5] and N. Gupta et al.[6] classify customers' emotions in customer support service and predict an appropriate emotions of agent for previous customers' emotion to improve customer experience. They used SVM and SVM-HMM to classify and predict emotions. Compare to baseline, which used only word features, features with previous emotions in dialogues helps classifying emotions.

Y. Wang et al.[7] detect emotion in social media. they employ several constraints, topic correlation, emotion bindings, and noisy labels, to detect emotions in short, noisy post in social media. They use Non-negative Matrix Factorizatioin

(NMF) to analyze sentiment of social media posts. it doesn't need learning phase, but need lots of information such as topic similarity between all pairs of documents, emotion binding probability, and feature document matrix.

III. Data

A. Data Collection

In order to find effectiveness of emotion features while classifying news articles, we select 10 world popular issues that could invoke users' negative, mixed or neutral, positive emotions respectively which is shown in table I. we collect news articles and users' posts which are related to the 10 issues on Facebook. Specifically, news articles are obtained from 4 famous news pages, CNN, BBC, Fox news, and Newyork Times whose followers are over 10 millions that is large enough to represent the Facebook users.

Since many people don't mention news articles in their posts, it's hard to obtain users' posts which are labeled with the news articles, so we use comments data below the news articles instead of posts. There are two reasons why posts can be replaced with comments, they both consist of several sentences and include users' thoughts and emotions. In addition, comments are all labeled with their news articles. we use three information from comments: date the comments was created, emotion extracted from comment, and comment itself as a feature vector labeled with issue on their article post.

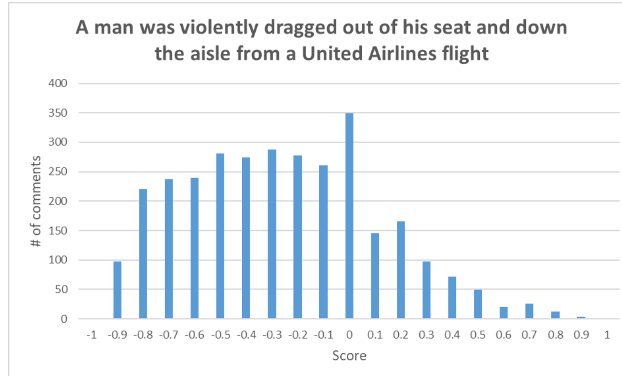
B. Data Statistics

We analyze the emotion of all comments data using Google Cloud Natural Language API[2]. The API analyzes them and calculates score and magnitude that score means emotion polarity of comment and magnitude means strength of emotion which are shown in Table II.

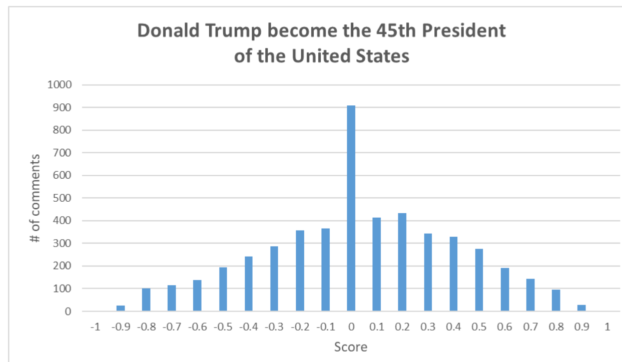
First, in order to find relationship that positive issue generates positive comments and vise versa, we analyze three issues and their comments which are negative, neutral, and positive respectively. Figure 4 shows the number of comments per score in positive, neutral, and negative issues. As you can see in Figure 4a, which is negative issue about United Airlines forcibly dragged out their passenger, the graph is skewed to negative side. It means, most people express negative feelings to the issue. On the other hand, in Figure 4c, which is positive issue about a woman travel every country on Earth as a woman, the graph is skewed to positive side. In addition, when the issue is neutral which is about election of United States, the graph is symmetric. It can be also seen in more detail on Table I.

As a result of analyzing 10 issues, it was found that the number of comment of 0 score is the highest in all data. There are several reasons, but the first reason is because data is obtained from Facebook. Many people wants to share issues with thier friends, so they "tag" their friends. Most tag comments only contain their friends' name, so that they have no emotion. Therefore, most comments have 0 score. Next, there are too many useless comments such as "lol", "wow", "LMFAO", and so on. These also have emotions, but they are not "Issue-Related" comments that they don't contains any information about issues. We eliminate very short comments which contains friends names or short exclamations. After eliminating, the number of training data is shown in Table I.

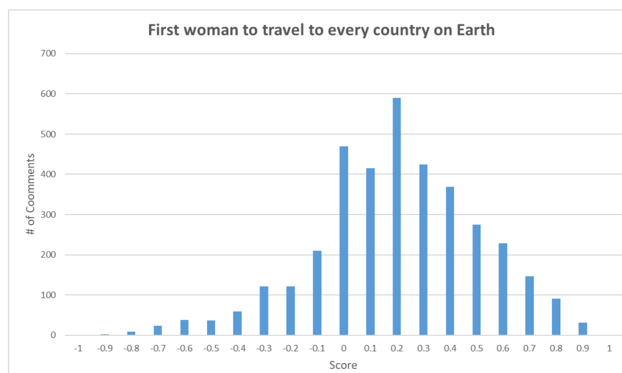
[2] https://cloud.google.com/natural-language/



(a) Negative News Article



(b) Neutral News Article



(c) Positive News Article

Fig. 4: The Number of Comments per Emotion Score of Each News Article

IV. Method

A. Emotion Extraction

In order to extract emotion from comments of social media, we use Google Cloud Natural Language API. Sentiment analysis of this API observes and identifies the dominant emotion and its magnitude from the given text. The emotion of text is depicted as a real number and it can be divided into three

TABLE I: Positive and Negative Comments on Each News Article

Issue Title	Negative	Neutral & Mixed	Positive	Total
Donald Trump become the 45th President of the United States	242(5.2%)	3338(71.8%)	1067(23%)	4647
First woman to travel to every country on Earth	34 (1%)	2062(63.6%)	1142(35.4%)	3238
Pope Francis says it's better to be an atheist than a greedy Christian	174(5.2%)	2398(71.9%)	759(22.9%)	3331
Man fatally shoots teens with AR-15 after they broke into his home	129(3.9%)	2683(81.2%)	490(14.9%)	3302
A man was violently dragged out of his seat and down the aisle from a United Airlines flight	556(18.4%)	2281(75.5%)	182(6.1%)	3019
Trump wants South Korea to pay for $1B missile defense system	55(5.3%)	871(85%)	99(9.7%)	1025
Thousands of protesters marching toward the White House for action on climate change	217(5%)	3329(77.2%)	764(17.8%)	4310
Emmanuel Macron's victory in the French presidential election	47(5%)	659(68.7%)	253(26.3%)	959
Harrowing security footage shows a woman fighting off car thieves by jumping on the hood of her car	275(6.1%)	3672(82%)	527(11.9%)	4474
A man went on a racist tirade after hearing another man speak Spanish on the phone	511(12.8%)	3194(80.1%)	281(7.1%)	3986

different classes: negative range from -1.0 to -0.75, neutral range from -0.75 to 0.25 and positive range from 0.25 to 1.0. The magnitude represents the overall strength of emotion from the given text which has range from 0.0 to infinity.

Table II shows an example of emotion extraction. Clearly positive emotion has positive number of score and high magnitude. Clearly negative emotion also has negative number of score and high magnitude. Neutral emotion has both score and magnitude near to zero. Mixed emotion between positive and negative has a score near to zero, but high magnitude. This is because, in the mixed emotion case, both high positive and negative scores cancel each out and have score around 0.0 with high magnitude of emotion.

TABLE II: Example of Emotion Extraction

Sentiment	Sample Values
Clearly Positive	"Score": 0.8, "Magnitude": 3.0
Clearly Negative	"Score": -0.6, "Magnitude": 4.0
Neutral	"Score": 0.1, "Magnitude": 0.0
Mixed	"Score": 0.0, "Magnitude": 4.0

B. Word Embedding

We map the text of comments to vectors of real numbers which is called word embedding [8]. Each words in the texts are projected into experimentally selected fixed length vector space. By using word embedding, it boosts the performance of learning speed for classification model. The dimension of vector space we use is 32 and the range of each real number is between -1.0 and 1.0.

C. Classification

We use deep neural network(DNN) model to classify the text of comments into matching news article. DNN model has been shown high performance in natural language processing task and also it makes easy to extend to integrate different types of data. As we use different types of data, such as comment, emotion and date, DNN model is beneficial for our task. The architecture of DNN is shown in Figure 5. Input data is combination among comment, emotion and date, and it turns positive integers into dense vectors of fixed size in embedding input layer. The DNN model has two hidden layers with ReLU activation function and each hidden layer has 100 units. We used batch normalization and dropout technique for regularization purpose. Output layer has ten units with softmax activation function since there are ten classes of social issue.

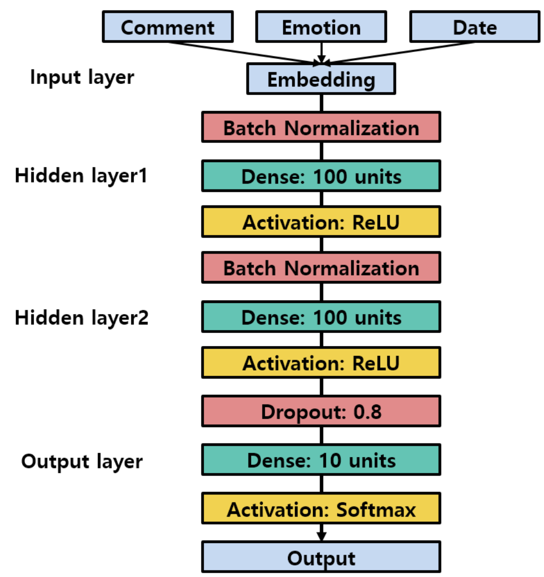


Fig. 5: Architecture of Deep Neural Network

V. EXPERIMENT

A. Experimental Setting

The acceptability of model is compared through the four different models: Comment only (baseline), Comment + Emotion (C+E), Comment + Date (C+D) and Comment + Date + Emotion (C+D+E) model. For training of DNN model, the dataset is divided into train, validation and test set, each respectively by 8 : 1 : 1 and the DNN model is trained through 100 epochs in Titan X GPU. Test set are used only once after the training is over.

B. Quantitative Evaluation

As shown in Table III, the models with emotion feature, such as C+E and C+D+E model, shows better performance.

And the models with date feature, such as C+D and C+D+E model boost performance significantly. However, as there are only ten news articles for the experiment, the date feature becomes highly dependent on each news article. So, the date model shows high performance. The important point on this experiment is that the model with emotion feature, such as C+E and C+D+E model, increase the performance on each experiment. It suggests that emotion feature is helpful to find out the matching news articles as each news articles have its own emotion distribution.

TABLE III: Test Accuracy on Each Model

	Comment (Baseline)	Comment + Emotion (C+E)	Comment + Date (C+D)	Comment + Date + Emotion (C+D+E)
Test Accuracy	0.668	0.690	0.982	0.986

C. Qualitative analysis

We conduct an analysis based on 58 randomly selected test data samples that shown in Table IV. Excluding the result that C+E and baseline model are both correct or wrong at the same time, C+E model win six to four comparing to the baseline model.

TABLE IV: Cross Comparison Between Baseline and C+E Model

	C+E Model is Correct	C+E Model is Wrong
Baseline is Correct	35	4
Baseline is Wrong	6	13

Table V shows the result that baseline model wins. In the second example, the comment “I wonder who he voted for”, for the following issue about “a man that went on a racist tirade” is a sarcasm by using the metaphor of Trump. We find that if the comment contains high level of pun or language play, then the model could not catch the relationship. In fact, the comments include a metaphor of Trump, so our model predicts well that the comment are related to news articles about “Trump become the president”. In our interpretation, this is because the apparent emotion and the actual emotion are different in the example.

Table VI shows the case of the result that C+E model wins. In the first example with the comment “I wouldn’t let it go either, it’s mine, I worked for it, I’d put up a fight, mad props”, “mad props” means extreme respect and agreement, for the following issue which is about “a woman fought off the car thieve”. You can see the commenter shows his or her respect by the direct consent. From this example, our model is able to catch the relationship between comments and related news articles, since emotion appeared on the comment can be used as a clue to find related news articles.

VI. Conclusion

In this paper, we propose three models to find a relative news article of posts in social media, especially in Facebook.

TABLE V: Case 1: Baseline Model Win

Issue	Comment	Remarks	C+E Output
Emmanuel Macron’s victory in the French presidential election	So glad Hillary Clinton did not win	Ironic humor	Climate Change
A man went on a racist tirade after hearing another man speak Spanish on the phone	I wonder who he voted for	Sarcasm	Donald Trump become the 45th President
Harrowing security footage shows a woman fighting off car thieves by jumping on the hood of her car	Poor thing. She did all that work and it looks like he still got away with the most important thing... her purse. hashtag lessonlearned	Metaphoric humor	Donald Trump become the 45th President

TABLE VI: Case 2: Comment + Emotion Model Win

Issue	Comment	Remark
Harrowing security footage shows a woman fighting off car thieves by jumping on the hood of her car	I wouldn’t let it go either, it’s mine, I worked for it, I’d put up a fight, mad props	Direct consent (Positive)
Watch Donald Trump become the 45th President of the United States.	How sad all you haters! You should be ashamed of yourselves. We thought the same of Obama but did not act like you are! You need to suck it up and keep your mouths shut like we did!	Anger outburst (Negative)
Thousands of protesters marching toward the White House for action on climate change	If just half of these people that are marching were doing something worthwhile like ¡unk¿ their time to the VA nursing homes ¡unk¿ police departments fire departments so many organizations homeless ¡unk¿ they all could use the volunteers give these people a feeling of ¡unk¿	Criticism (Negative)

In order to overcome the problem of keyword only model, we use emotion feature of each post. We have demonstrated that, each news article has own emotional distribution and it shows that emotion feature is a necessary feature for classification of news article. We show this with the quantitiative analysis, which classification model with emotion feature gives better performance than classification model without emotion feature. Also, in the qualitative analysis, we show classification model with emotion works better than without emotion when there is apparent emotion on the post.

For future work, we will generalize our model with various social media such as Twitter and Instagram. In addtion, we will experiment in practical environment where over 1000 news articles are generated in a day.

References

[1] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-based methods for sentiment analysis,” *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[2] B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,” in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, p. 271.

[3] M. Ghiassi, D. Zimbra, and S. Lee, "Targeted twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks," *Journal of Management Information Systems*, vol. 33, no. 4, pp. 1034–1058, 2016.

[4] S. Poria, H. Peng, A. Hussain, N. Howard, and E. Cambria, "Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis," *Neurocomputing*, 2017.

[5] J. Herzig, G. Feigenblat, M. Shmueli-Scheuer, D. Konopnicki, A. Rafaeli, D. Altman, and D. Spivak, "Classifying emotions in customer support dialogues in social media," in *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016, p. 64.

[6] N. Gupta, M. Gilbert, and G. D. Fabbrizio, "Emotion detection in email customer care," *Computational Intelligence*, vol. 29, no. 3, pp. 489–505, 2013.

[7] Y. Wang and A. Pal, "Detecting emotions in social media: A constrained optimization approach." in *IJCAI*, 2015, pp. 996–1002.

[8] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 1019–1027.