Classifying aggravation status of COVID-19 event from short-text using CNN

Ekasari Nugraheni
Research Center for Informatics
Indonesian Institute of Sciences
Bandung, Indonesia
ekasari.nugraheni@lipi.go.id

Purnomo Husnul Khotimah
Research Center for Informatics
Indonesian Institute of Sciences
Bandung, Indonesia
hkhotimah@mail.informatika.lipi.go.id

Andria Arisal
Research Center for Informatics
Indonesian Institute of Sciences
Bandung, Indonesia
andria.arisal@lipi.go.id

Andri Fachrur Rozie
Research Center for Informatics
Indonesian Institute of Sciences
Bandung, Indonesia
andri@mail.informatika.lipi.go.id

Dianadewi Riswantini
Research Center for Informatics
Indonesian Institute of Sciences
Bandung, Indonesia
dianadewi@mail.informatika.lipi.go.id

Ayu Purwarianti
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
ayu@std.stei.itb.ac.id

***Abstract*—COVID-19 pandemic is a new precedent that has changed many aspects of human life. With the uncertainty of vaccine availability, stakeholders are required to track the dynamics of COVID-19 events to prepare the necessary response. One sub-task in tracking the dynamics of an event is to identify the aggravation status of the event (i.e., whether an event is worsening or getting better). We experimented with convolutional neural network (CNN) models to classify the status of COVID-19 aggravation status from a short text. CNN without one hot encoding prevailed. Furthermore, we conduct tuning to achieve better performance of CNN. The highest performance was achieved by tuning some of the configuration parameters. As the final result, the model performed at best (accuracy = 87.585% and F1-score = 76%) when using 80 nodes, SGD optimizer, lr = 0.1, and momentum = 0.9.**

***Index Terms*—short text, news title, CNN, covid, sentiment analysis**

I. Introduction

Coronavirus disease 2019 (COVID-19) began in Wuhan, China, in December 2019, declared by the World Health Organization (WHO) as the Public Health Emergency of International Concern on 30 January 2020 and as a pandemic on 11 March 2020 [1].

With the uncertainty of vaccine availability in the near future, stakeholders (such as government, health institutions, societies, etc.) need to be aware of the next wave. Tracking the dynamics of COVID-19 events is essential for stakeholders so that necessary responses could be prepared in advance. A sub-task in tracking the dynamics of the event is to identify its aggravation status, a status that describes whether an event is worsening or improving. The official reports are often used to analyze the aggravation status of COVID-19 events. However, official reports on COVID-19 are frequently slow due to the hierarchical nature of its reporting system [2] [3].

The implementation and use of web technologies generate abundant data every day. One of them is the online news media, which reports on various events, including cases of COVID-19. Thus, online news media has a great potential as an informal but reliable data source for monitoring COVID-19 dynamics in near to real-time. However, analyzing the contents of the whole news will be heavy in computation. Another option is by using the news title. Even though news titles are short, the main topic of the news is commonly represented in the title text. For example, *“Lockdown” dilonggarkan, kasus corona kembali meningkat di Jerman* (in English: “ Lockdown ” is relaxed, corona cases are back on the rise in Germany) is a news title in Indonesian language. From this title, we are able to understand that the situation of corona cases aggravates in Germany. Another title is *3 Hari Berturut-turut Tak Ada Tambahan Kasus Baru Covid-19 di Denpasar* (in English: 3 Consecutive Days No Additional New Covid-19 Cases in Denpasar). Again, from this title, we are able to understand that the situation in Denpasar is improving. Based on these two examples, we believe that information about COVID-19 situation changes (dynamics) is able to be harnessed from news titles.

Our problem falls into a text classification problem that is similar to sentiment analysis. Sentiment analysis is the field of study that analyzes people’s opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in a written text [4].

Many studies had been conducted in the field of sentiment analysis. [5] identified that some researchers focused on target-dependent sentiment classification, which is to classify sentiment polarity for a given target on sentences consisting of explicit sentiment targets. For example, to identify sentiments towards some products (target) from user opinion in twitter dataset [6]. Furthermore, other researchers dealt with mining opinions in more specialized sentence structure, such as comparative sentences, to determine the degree of positivity surrounding the analysis of comparative sentences [5]. As an example, [7] mined sentiment's of opinion in a typical comparative sentence compares two or more entities. There has also been work focusing on sentiment analysis of conditional sentence structure [8], or sentences with modality [9], which have some special characteristics that make it hard for a system to determine sentiment orientations. Nevertheless, the aforementioned studies focus on appraising the subjectivity of a text that is the opinion expressed in the text. As sentence can be classified into subjective (opinions) and objective (facts) [10] [11], our study focus in classifying objective text that is news title containing facts surrounding the event of COVID-19, to identify the situation of an event whether the situation is aggravating or not. The study will focus on online news with Indonesian language because many of the work on filtering relevant events from text or done for text in English.

In this study, we experimented with CNN models to address the problem defined. We investigated CNN's performance when using one hot encoding and when not using one hot encoding. The model with the best performance was further tuned to attain better results. The remaining of the paper is structured as follows: Related work is described in section II, the framework of our method is explained in section III, and the results and discussion of our experiment is presented in section IV. We conclude the paper in section V.

II. Related work

When a predefined monitored event is detected, it is important to identify the event's dynamics in order to have a better understanding of the situation. For example, [12] used twitter data to monitor natural disaster social dynamics. This study focused on the location identification in tweets to facilitate the reporting of trapped individuals, the provision of medical assistance, and the delivery of basic needs, such as food, water, and shelter. Another study is in the medical field, [13] and [14] used medical notes to gain insight towards patients' clinical outcomes by analyzing the sentiment of subjective expressions made by clinicians. However, both studies focus on using sentiment analysis for prediction purposes.

A study that is close to our work is done by HealthMap in measuring infectious diseases activity. Similar to our study, HealthMap uses online news as one of their data sources and provides a heat index to show hot spots and general levels of outbreak activity. The "heat" rating is based on the number of alerts, diseases, and feeds for each day at each location to show the general levels of outbreak activity [15]. However, news volume only shows the scale of the event. It does not show the real dynamics of the event, such as whether the event is worsening or getting better. Providing activity measurement based on such data should be able to provide a finer grain information for the stakeholders.

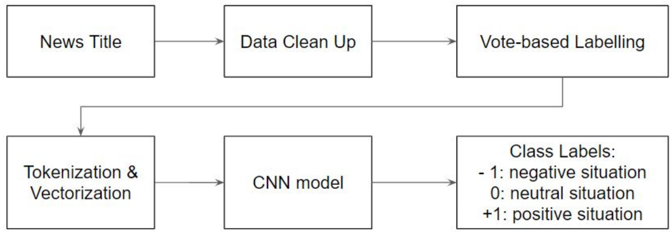

Fig. 1. Experiment framework for classifying COVID-19 situation from news title.

Compared to social media text, such as tweets, news text is considered to have higher credibility and lower noise [16]. In this study, we used the title part of the news as the data source that is considered as a short text. A similar study using short text form news text is [17]. Balahur, et al. [17] used quotation text extracted from news content to determine the sentiment of the entities found in the news. Their approach is different from our objective, which is to use sentiment analysis to capture the aggravation status of a predefined event that is COVID-19.

For the method, we conducted a supervised learning method by using a convolutional neural network (CNN, or ConvNet) models. CNN has attracted high attention in several recent studies in natural language processing. A CNN [18] is a class of deep neural network architecture which most commonly applied for visual imagery analysis. However, CNN models have shown to be effective and give a good metrics performance for various tasks in NLP [19]. In addition, ConvNets can be directly applied to distributed or discrete embedding of words, without any knowledge on the syntactic or semantic structures of a language [20]. This advantage makes CNN a practical method compared to classical NLP methods.

III. Methodology

We conducted the experiment using the framework shown in Fig. 1. Collected news titles are cleaned up and labeled using a votes-based mechanism. The cleaned and labeled new titles are tokenized and vectorized to be transformed into matrix input for the model. We used multi categorical output (three class labels) to classify the situation of COVID-19 events from the news titles.

A. Dataset

The data was acquired by conducting web crawling, a procedure for browsing online news sites on the internet, and downloading the content it finds for further storage in a database. Online news data collections were collected from Indonesian online news portals, including `detik.com`, `kompas.com`, `antara.com`, and `merdeka.com`. 20,431 news titles were obtained from the crawling using 'corona' and 'covid' as the keywords.

The pre-processing process applied to raw data (20,431 data) is straightforward because the news title is already using formal sentence structure. We only removed the HTML tags

TABLE I
EXAMPLES OF NEWS TITLE FOR EACH DATA CLASS

Class Label	Title
negative	*1 Kampung di Garut Diisolasi Usai Ditetapkan Zona Merah Corona* (Eng. title: 1 Village in Garut Isolated After Establishing the Corona Red Zone.) *1 Napi Positif Corona , 194 Napi di Lapas Bojonegoro Jalani Tes dan 6 Dikarantina* (Eng. title: 1 Positive Corona Convict, 194 Prisoners in Bojonegoro Prison Underwent Tests and 6 Quarantined.) *106 Orang Meninggal, 16 Negara Ini Konfirmasi Terinfeksi Virus Corona* (Eng. title: 106 People Died, These 16 Countries Confirm They Are Infected With Corona Virus)
neutral	*11 mitos yang dianggap dapat mencegah virus corona* (title: 11 myths that are thought to prevent the corona virus.) *135 Napi Asimilasi Covid-19 Kembali Dijeruji* (Eng. title: 135 Covid-19 Assimilation Prisoners Are Again Tested.) *Ahli Kandungan: Covid-19 Tidak Bisa Menular ke Bayi Dalam Kandungan* (Eng. title: Obstetrician: Covid-19 Cannot Be Transmitted to Infants in the Womb)
positive	*10 Pasien Sembuh Covid-19 di Surabaya Disambut Suka Cita* (Eng. title: 10 Patients Healing Covid-19 in Surabaya are Welcomed with Love.) *10 Ribu Pasien Corona Singapura Sembuh, China Tanggapi Ancaman Trump ke WHO.* (Eng. 10 Thousand Corona Singapore Patients Recover, China Responds to Trump's Threat to WHO.) *125 warga Jakarta Pusat sembuh dari COVID-19* (Eng. title: 125 residents of Central Jakarta recovered from COVID-19)

and duplicate data. Data duplication may happen when the news is divided into several pages. The number of the dataset after going through this process is 16,833 data.

B. Vote-based labelling

Vote-based labeling is done to ensure the quality of the labeling process. The process is carried out by three respondents who labeled the news titles based on its aggravation status, and then the title's label was decided based on the most votes. We divided the sentiments that represent the aggravation situation of COVID-19 events that might be shown in the news title into three classes:

- -1 (negative): news of the COVID-19 incident that reports negative events that could potentially aggravate the situation; such as, reports about the number of patients has increased, people infected with COVID-19, positive test result, and worsening zone discoloration of the area.

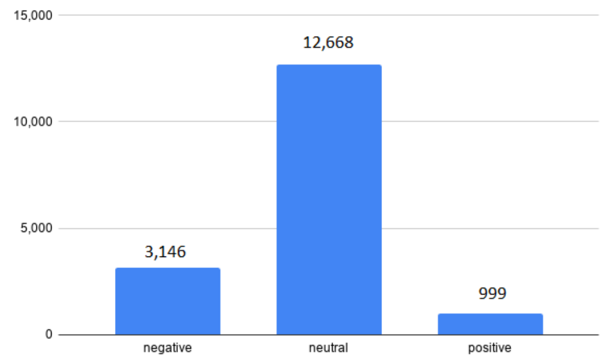

Fig. 2. Multicategorical sentiment data

- 0 (neutral): news of the COVID-19 incident that does not show any indication of positive or negative events, for example, a report that a rapid test was carried out and the results are not yet known, donations for COVID, and news headlines that are not related to corona.
- +1 (positive): news of the COVID-19 incident containing positive (hopeful) events that show the situation is improving; such as, information that patients are recovering, less COVID-19 cases, and negative test results.

Table I shows examples of news titles for each data class. As for the labeling result, it is shown in Fig. 4. As shown in Fig. 4, we acquired an imbalanced dataset. The dataset is dominated by neutral data (12,668). Negative and positive data numbers are 3,146 and 999, respectively.

C. CNN model

Even though it was developed for image processing, the convolution neural network (CNN) can also be applied for natural language classification by applying the convolution in one dimension instead of two-dimension on image classification. In text analysis, CNN uses a series of window filters to create a hierarchical simpler pattern from a more complex input pattern. The model is built by creating an input matrix after loading and tokenizing the dataset. In our experiment, we treat the input matrix with and without one-hot encoding. The input is then passed to the embedding layer. The output of the embedding layer is passed to a convolution 1D layer. Within this layer, we experimented with various hyperparameters to find an optimal configuration. We pass the output of the convolution layer to a max pooling layer, which reduces the tensor dimension. The tensor is then passed through a flatten layer to reshape into features vector, which passes to a densely connected layer to merge the neurons using Relu and Softmax activation functions. The detail of this architecture is shown in Fig. 3.

IV. RESULTS AND DISCUSSION

Our web crawler collected 16,833 news titles from various online portals in Indonesia. We split 13,116 of data as training data and the other 3,717 as testing data. Other than accuracy,

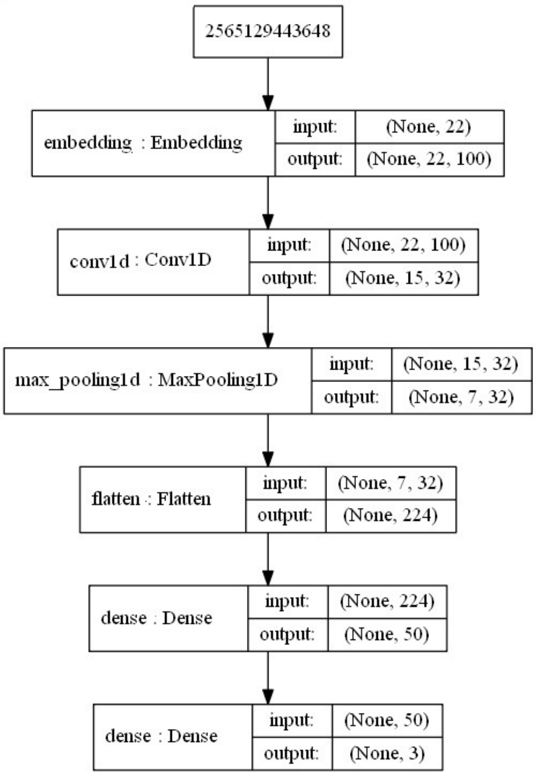

Fig. 3. CNN Architecture

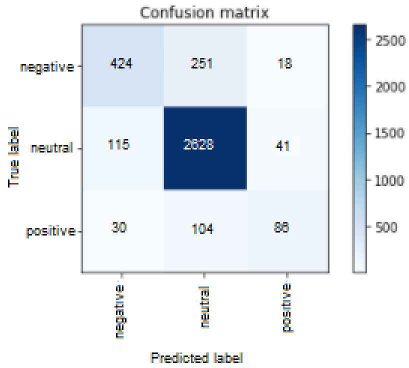

Fig. 4. Confusion matrix before tuning.

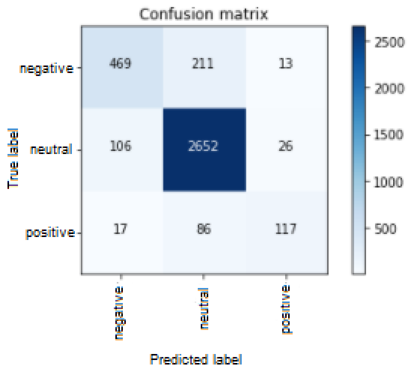

Fig. 5. Confusion matrix after tuning.

we also use F1-score because of the imbalance class dataset and training duration as other comparison metrics of built models. In the initial experiment, we used two models, namely:

- CNN+Embedding layer with one-hot encoding (model 1)
- CNN+Embeding layer without one-hot encoding (model 2)

Training and testing were carried out using variants of epochs 10,50, and 100. The results of the initial experiment are shown in Table II, where the highest accuracy and f1-score value are achieved by model 2 (CNN+Embedding layer without one-hot encoding) with epochs = 10. Therefore, further experiments will be carried out using this model.

In machine learning, parameter configuration can not be instantly specified for a model. The configuration of each parameter must be obtained by conducting training and testing using different parameter values. Hence, we conducted parameters fine tuning for the second experiment. Based on [21], some parameters could improve or accelerate the adaptation of the model weights in response to a training dataset which includes:

- epochs
- batch size which controls the number of training samples processed
- the number of nodes and layers to control the model capacity
- loss function to calculate model error
- optimizer
- learning rate (lr) to control how fast the model will learn the problem
- the momentum of the learning rate, which has the effect of smoothing the optimization process

To get better performance, we configured the CNN+Embedding model's parameters listed before. As the first step, we varied the number of epochs, batch size, and layer; and as a result, the model achieved the best performance at epochs = 10, batch size= 32, and layer = 1. Moreover as the second step, we varied the number of nodes (50 nodes and 80 nodes), optimizer (Adam and SGD), and loss function (categorical cross-entropy and Kullback leibler divergence). Sparse categorical cross-entropy loss-function was not used because this loss function is commonly used for a high number of classes. Additionally, SGD's learning rate and momentum were varied with 0.1 and 0.01 for the learning rate; and 0.7 and 0.9 for the momentum value. The results of this experiment are shown in Table III and Table IV.

From this experiment, we found that by using Adam optimizer, the model is required to perform a significantly more training time compared to SGD optimizer in both nodes configuration (50 nodes, 80 nodes). However, SGD optimizer needs to be fine tuned to surpass Adam optimizer's performance. Furthermore, when using categorical cross-entropy as the loss function, the model performed better in general when

TABLE II
METRICS COMPARISON ACCURACY, F1-SCORE AND TRAINING TIME

	CNN+Embedding(one_hot)			CNN+Embedding(without one-hot)		
epochs	Accuracy%	F1-score%	time (ms)	Accuracy%	F1-score%	time (ms)
10	83.338	67.33	49,397.850	84.393	68.33	44,738.910
50	82.743	66.66	215,538.374	83.420	67.66	183,049.404
100	82.743	66.00	415,346.269	83.013	67.33	407,274.888

TABLE III
CONFIGURATION TUNING ON MODEL CNN+EMBEDDING WITHOUT ONE-HOT ENCODING (50 NODE, EPOCHS 10, BATCH SIZE 32, 1 LAYER)

Loss Function	categorical crossentropy			Kullback leibler divergence		
Optimizer	Accuracy%	F1-score%	time (ms)	Accuracy%	F1-score%	time (ms)
Adam	84.393	68.33	44,738.910	85.475	71.00	41,946.1579
SGD(lr=0.01, Momentum=0.7)	81.634	48,33	30,832.264	81.309	46.66	29465.5876
SGD(lr=0.01, Momentum=0.9)	84.636	70.33	30,662.801	85.502	71.66	29,951.7185
SGD(lr=0.1, Momentum=0.7)	86.935	73.33	29,304.235	86.530	73.66	29,016.1385
SGD(lr=0.1, Momentum=0.9)	85.475	72.00	30,616.047	86.692	74.00	29,950.8543

TABLE IV
CONFIGURATION TUNING ON MODEL CNN+EMBEDDING WITHOUT ONE-HOT ENCODING (80 NODE, EPOCHS 10, BATCH SIZE 32, 1 LAYER)

Loss Function	categorical crossentropy			Kullback leibler divergence		
Optimizer	Accuracy%	F1-score%	time (ms)	Accuracy%	F1-score%	time (ms)
Adam	84.998	70.33	39,959.737	84.744	69.66	42,130.312
SGD(lr=0.01, Momentum=0.7)	81.715	47.66	30,832.264	83.094	49.66	29,157.948
SGD(lr=0.01, Momentum=0.9)	84.717	69.00	35,673.588	84.907	70.33	31,812.415
SGD(lr=0.1, Momentum=0.7)	86.070	73.00	31,221.057	85.231	71.66	31,467.522
SGD(lr=0.1, Momentum=0.9)	87.585	76.00	30,616.047	85.231	69.33	30,410.247

using fewer nodes (80 nodes). This condition contrasts with the model using Kullback leibler divergence as a loss function, which performs better in more nodes (50 nodes). As the final result, the model performed at best (F1-score 76% and 87.585%) using 80 nodes, SGD optimizer, lr = 0.1, and momentum = 0.9, as shown in Table IV.

To show the improvement of the model after the fine tuning, Fig. 4 and Fig. 5 shows the confusion matrix before and after tuning, respectively. From the figures, we are able to see that the number of true positives increases.

V. CONCLUSION

We have presented a neural network approach to classify aggravation status of COVID-19 events from news titles. We classified the news situation expressed in the news titles into three categories; negative (i.e., situations that indicate the aggravation status is worsening), neutral (i.e., situations that do not show any indication whether there is a positive or negative case of COVID-19), and positive (i.e., situations that indicate COVID19 event is getting better). To the best of our knowledge, little studies have been done in capturing the situation status of an event from news titles. Additionally, we presented the effects of parameters tuning towards the models' performances. Our initial experiment showed that, without any customization, CNN performed better without using one-hot encoding for our task. With further configuration tuning, we were able to increase the accuracy by 3.192% (from 84.393 % to 87.585 %) and F1-score by 7.67% (from 68.33% to 76.00%).

VI. DECLARATION

Author contribution

Purnomo Husnul Khotimah and Ekasari Nugraheni contributed equally as main authors. Andria Arisal, Andri Fachrur Rozie, Dianadewi Riswantini, and Ayu Purwarianti contributed equally as the member contributors of this paper. All authors read and approved the final paper.

Conflict of interest

The authors declare no conflict of interest.

REFERENCES

[1] WHO. Coronavirus disease (covid-2019) situation reports. world health organization.

[2] L. Madoff, "Infectious diseases surveillance and alert systems," *International Journal of Infectious Diseases*, vol. 16, p. e45, 2012.

[3] Y. T. Yang, M. Horneffer, and N. DiLisio, "Mining social media and web searches for disease detection," *Journal of public health research*, vol. 2, no. 1, p. 17, 2013.

[4] J. Zhao, K. Liu, and L. Xu, "Sentiment analysis: mining opinions, sentiments, and emotions," 2016.

[5] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using bilstm-crf and cnn," *Expert Systems with Applications*, vol. 72, pp. 221–230, 2017.

[6] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent twitter sentiment classification," in *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, 2014, pp. 49–54.

[7] M. Ganapathibhotla and B. Liu, "Mining opinions in comparative sentences," in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 2008, pp. 241–248.

[8] R. Narayanan, B. Liu, and A. Choudhary, "Sentiment analysis of conditional sentences," in *Proceedings of the 2009 conference on empirical methods in natural language processing*, 2009, pp. 180–189.

[9] Y. Liu, X. Yu, Z. Chen, and B. Liu, "Sentiment analysis of sentences with modalities," in *Proceedings of the 2013 international workshop on Mining unstructured big data using natural language processing*, 2013, pp. 39–44.

[10] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, pp. 129–136.

[11] J. Wiebe and T. Wilson, "Learning to disambiguate potentially subjective expressions," in *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.

[12] W. Wan, B. Liu, Z. Zeng, X. Chen, G. Wu, L. Xu, X. Chen, and Y. Hong, "Using cygnss data to monitor china's flood inundation during typhoon and extreme precipitation events in 2017," *Remote Sensing*, vol. 11, no. 7, p. 854, 2019.

[13] T. H. McCoy, V. M. Castro, A. Cagan, A. M. Roberson, I. S. Kohane, and R. H. Perlis, "Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: an electronic health record study," *PloS one*, vol. 10, no. 8, p. e0136341, 2015.

[14] G. E. Weissman, L. H. Ungar, M. O. Harhay, K. R. Courtright, and S. D. Halpern, "Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness," *Journal of biomedical informatics*, vol. 89, pp. 114–121, 2019.

[15] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, "Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports," *Journal of the American Medical Informatics Association*, vol. 15, no. 2, pp. 150–157, 2008.

[16] P. Kostkova, "A roadmap to integrated digital public health surveillance: the vision and the challenges," in *Proceedings of the 22nd International Conference on World Wide Web*, 2013, pp. 687–694.

[17] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. van der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment analysis in the news," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[20] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649–657.

[21] J. Brownlee, *Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions*. Machine Learning Mastery, 2018.