Does it make you sad? A lexicon-based sentiment analysis on COVID-19 news tweets

Dedy Suryadi
Industrial Engineering Department, Parahyangan Catholic University
Jl. Ciumbuleuit No. 94, Bandung 40141, Indonesia
E-mail: dedy@unpar.ac.id

Abstract. This research utilizes a lexicon-based sentiment analysis to reveal the emotions conveyed by various news media and analyze the differences among the media. NRC Affect Intensity Lexicon is utilized; it provides the emotions of words in 8 categories of emotions. The lexicon is modified to reflect the context of the news, i.e., COVID-19 pandemic. Tweets are collected from four Indonesia’s news media and three international English-language news media. Each tweet from a news media is assigned a total emotion score for each of the 8 emotions. The scores are averaged for each day to obtain the Daily Emotion. Based on the Daily Emotion, the Dominant Daily Emotion may be identified, i.e., the emotion with the highest average score in a particular date. The Dominant Daily Emotion visualization shows that the dominant emotions in Indonesian-language media are Sadness and Trust, while Fear and Trust are considerably more pronounced in English-language media. Furthermore, among Indonesian-language media, kompascom is significantly more intense than the others in Joy, Sadness, Fear, and Trust. Among English-language media, XHNews is more intense than others in conveying Fear, Sadness, and Trust; while XHNews and timesofindia are both the most intense in Joy. The sentiment analysis in this research concludes that the Emotion Mix of news media are significantly different. Furthermore, news media also convey significantly different levels of emotion intensity in reporting COVID-19 news.

1. Introduction
Twitter has been one of the main sources of information nowadays. News media have been actively utilizing Twitter as well to broadcast information, including the news about the COVID-19 pandemic. Due to the limitation of number of characters, news media typically tweet concise news or striking headlines.

Previous research has shown that the news content triggers different types of emotion. Lin et al. [1] predict the emotions of Useful, Bored, Happy, Angry from the news in Yahoo Chinese with higher than 73% accuracy. Furthermore, Zhu et al. [2] model emotions with respect to time, such that the evolution of the emotions may be tracked.

More specifically, Kim and Cameron [3] reveal that news frames (e.g., sadness-inducing) induce distinct emotions. In other words, different framing of the same news may affect the degree of message processing by the readers [4], as well as the readers’ emotional response, i.e., media influence what people feel. Boukes and Vliegenthart [5] conclude that the consumption of news that focus on misery, powerlessness, and negativity negatively affect the mental well-being. Therefore, it is clear that the emotions in the news, including news tweets, do affect the readers. Thus, the first objective of this
research is to analyze the emotions that are conveyed by the news tweets from various news media. The emotion analysis in this paper takes the approach of a lexicon-based sentiment analysis.

From the perspective of news media, emotional communication is indeed one of journalists’ main strategies in writing news, according to a research by Saxer and Märki-Koepp [6]. Negativity, in particular, is one of journalists’ key news values [7]. Therefore, this research’s second objective is to analyze the differences in the emotional communication among news media, e.g., whether or not some news media conveys significantly higher sadness than others.

2. Literature review: Sentiment analysis
Sentiment analysis is defined as a computational study of sentiments, emotions, opinions, as well as attitude, towards a particular aspect in texts [8]. A typical sentiment analysis task is determining whether a sentence expresses a positive or negative sentiment towards an aspect [9]. According to Ravi and Ravi [9], sentiment analysis may be performed using machine learning, lexicon-based, or hybrid approach. A machine learning approach generally requires annotated sentences to train the machine learning model and the model is subsequently applied to classify the unlabeled sentences. A lexicon-based approach, on the other hand, relies on the sentiment lexicon that has been developed from crowdsourcing or similar method.

Sentiment lexicons may have different number of categories of emotion. In NRC Affect Intensity Lexicon [10], there are 8 emotion categories. The categories correspond to the basic emotions that are defined by Plutchik [12]. Plutchik’s basic emotions are conceptualized in terms of pairs of opposites, i.e., Joy-Sadness, Trust-Disgust, Fear-Anger, and Surprise-Anticipation.

The NRC Affect Intensity Lexicon is chosen in this research because it provides a wider range of emotions to analyze the news tweets, compared to using binary sentiment lexicon, i.e., positive and negative sentiments only. Furthermore, the lexicon also provides real-valued scores for rating the intensity of the 8 basic emotions. The lexicon is built using Best Worst Scaling technique, in which annotators pick the best and the worst items among $n$ items regarding to the property of interest [11], e.g., a particular emotion. Based on the selections, the real-valued emotion intensity score of a word may be inferred.

3. Methodology, data and results
3.1. Data collection and preprocessing
The data in this paper is obtained from Twitter. In particular, the tweets are collected from the official Twitter accounts of several news media in several countries.

The news media from Indonesia and their respective verified Twitter accounts are as follows: The Jakarta Post @jakpost, CNN Indonesia @CNNIndonesia, Republika.co.id @republikaonline, Kompas.com @kompascom. Among those four news media, The Jakarta Post is the only news media that uses English, while the others use Indonesian. The other three news media whose tweets are collected are China’s China Xinhua News @XHNews, India’s The Times of India @timesofindia, and Singapore-based CNA @ChannelNewsAsia. All three news media use English. The Twitter account name will subsequently be used to refer to each news media.

The tweets are collected using a Python package called GetOldTweets3. The collection period is from January 1, 2020 to May 31, 2020. Tweets are collected if they contain either “corona” or “covid” (case insensitive). The highest number of tweets is obtained from kompascom (19,357) and the lowest is jakpost (2,032). An example of jakpost’s five last tweets during the data collection period are shown in Table 1. The tags and HTML links are subsequently removed from the tweets for further text processing.

The tweets are preprocessed as follows: replacing every capital letter with a lowercase, removing words that start with “http” and a hashtag (“#”), and removing any non-alphanumeric character at the beginning or end of a word. For the tweets in Indonesian language, a highly frequent bigram “rumah
“sakit” (hospital) is replaced by “rumah_sakit” and the abbreviations that refer to a hospital (i.e., “rs”, “rsud”, and “rsup”) are all replaced by “rumah_sakit”.

3.2. NRC Affect Intensity Lexicon Modification

The NRC Affect Intensity Lexicon [10] is utilized to score each tweet on each of the 8 emotions. The lexicon is also available for Indonesian language. Therefore, the tweets from Indonesian news media may be scored using the same lexicon.

| Timestamp       | Tweet                                                                 |
|-----------------|----------------------------------------------------------------------|
| 2020-05-31 13:10:44 | UN to use Liberian president Weah's COVID-19 awareness song #jakpost |
| 2020-05-31 12:21:08 | Via Vallen explained the story about her younger brother who has been tested positive for COVID-19. #jakpost |
| 2020-05-31 11:34:05 | Canada to promote holidays at home because of COVID-19 border closures #jakpost |
| 2020-05-31 11:04:47 | The Health Ministry announced 700 new confirmed COVID-19 cases on Sunday, bringing the total number of infections nationwide to 26,473. #jakpost |
| 2020-05-31 10:10:47 | COVID-19: Surabaya mayor feuds with East Java governor over mobile PCR labs #jakpost |

Nevertheless, several modifications must be made to the lexicon such that the emotion scores reflect the context of the tweets, i.e., COVID-19 pandemic. The modifications are subjective, yet deemed reasonable.

For the English lexicon, two modifications are made as follows:
1. Swapping: the emotions of “positive” and “negative” are swapped because, in the pandemic context, “positive” has a negative meaning.
2. Removing the words that do not convey emotions in the pandemic context, such as the names of days in a week; as well as some words that are considered containing no emotions when they appear in a phrase, i.e., “health” in “health ministry”, and “dream” in “dream world”.
3. Flipping: the emotions of “highest” are flipped to their corresponding opposites when it appears together with “case(s)” and “death(s)”. 

For the Indonesian lexicon, the modifications are made as follows:
1. Swapping: the emotions of “positif” (positive) an.d “negatif” (negative) are swapped for the same reason with the English lexicon.
2. Removing the words that, in the pandemic context, do not match the original English terms from which they got translated; such that these words are assigned zero score for all emotions. Those words are as follows:
   a. “sangat” (very; adverb) which is translated from “wonderfully” in the English lexicon, 
   b. “sedang” (be; verb) is translated from “temperate”,
   c. “tetap” (still; adverb / remain; verb) is translated from “fixed”,
   d. “jantung” (heart) –mostly refers to heart disease; is translated from “heart” with joy as one of the emotions,
   e. “khusus” (special) –mostly refers to the special spokesperson of COVID-19; is translated from “special” with joy as as one of the emotions;
   f. “kesehatan” (health) –mostly refers to the health ministry; is translated from “health” with joy as one of the emotions.
3. Adding: an entry “rumah_sakit” (hospital) is added to the lexicon and the emotions are copied from the English lexicon for the word “hospital”.
4. Flipping: the emotions of the following words, i.e., “meningkat” (increase), “mencapai” (reach), “melonjak” (soar), “jiwa” (lives), “ledakan” (explosion); are flipped when they appear together with “kasus” (case), “kematian” (death), “meninggal” (die).

The modifications are made based on the preliminary observations that some tweets were assigned high scores for joy, despite those tweets clearly being bad news. For example, before the modifications, tweets that mention “highest deaths” score high on joy. These modifications are not claimed to be exhaustive but they certainly help in eliminating the extreme cases, such as the aforementioned case of “highest deaths”.

3.3. Emotion scoring for tweets: total emotion, daily emotion, dominant daily emotion

Each tweet is split into words. If a word appears in the NRC Affect Intensity Lexicon, then its corresponding emotion(s) are collected. The total score for each emotion in a tweet is then obtained by summing the scores over all words related to the emotion. The procedure is illustrated in Table 2 with a tweet from jakpost on May 8, 2020. The tweet content is “Under normal circumstances, Buddhists would have gathered in temples across the country on Thursday to celebrate Waisak. But this year, with COVID-19 measures preventing most gatherings, the temples have gone quiet. #jakpost”.

| Word in Lexicon | Score (joy) | Score (trust) | Score (sadness) |
|-----------------|-------------|---------------|-----------------|
| circumstances   | 0.125       | 0.359         | 0               |
| celebrate       | 0.844       | 0             | 0               |
| gone            | 0           | 0             | 0.453           |
| quiet           | 0.188       | 0             | 0.094           |
| Total Emotion   | 1.157       | 0.359         | 0.547           |

Each day, a news media may broadcast several tweets. To reflect the daily emotion, the total emotions for every tweets on a particular date are averaged. An example of the daily emotion in jakpost on May 31, 2020, over 9 tweets on that particular date, is shown in Table 3. Based on the daily emotion, the dominant daily emotion may be obtained, i.e., the emotion with the highest score in the daily emotion.

Using the example in the Table 3, the dominant daily emotion for jakpost on May 31, 2020 is Trust.

| Anger | Anticipation | Disgust | Fear | Joy | Sadness | Surprise | Trust       | Score (dominant) |
|-------|--------------|---------|------|-----|---------|----------|-------------|------------------|
| 0     | 0.038222     | 0.073778| 0.142222| 0.152778| 0       | 0.506778  |             | 0.506778        |

The plots of daily emotion for different news media indicate the differences between media. For example, the plots in Figure 1 display the difference in daily Joy and Sadness emotions between kompascom and jakpost. Sadness (purple line) is obviously above Joy (orange line) in kompascom, while such dominance does not exist in jakpost. This observation motivates the further analysis whose results are presented in the next sections.

3.4. Emotion Mix of News Media

First of all, the emotion mix, i.e., the average proportions of the emotions in the daily emotion; of each news media is shown in Figure 2. The visualizations in this and subsequent subsections are produced by the Python package called Matplotlib [13].
Figure 1. Daily Emotions in kompascom (Left) and jakpost (Right).

On the left part of Figure 2, it can be observed that the emotion mix of Indonesian-language news media contains a significantly higher portion of Sadness (i.e., at least 29.5%) than jakpost. Similar to jakpost, the international English-language news media also contain just around 20% of Sadness in their emotion mixes.

Figure 2. Emotion Mix in Indonesia’s News Media (Left) and English-language News Media (Right).

On the other hand, the portion of Fear appears slightly higher in English-language news media (i.e., at least 31.1%) than their Indonesian-language counterparts, with ChannelNewsAsia as the highest with 36.8%. The portion of Sadness is just around 20% in these news media.

3.5. Dominant Daily Emotion of News Media
The four emotions are selected based on the most common dominant daily emotions, which are displayed in Figure 3 for Indonesia’s news media and Figure 4 for English-language news media. It is apparent that Indonesian-language news media’s preferred dominant emotion is Sadness, with CNNIndonesia as the leader with 72.6%; while kompascom’s dominant emotions are Trust and Sadness.

On the other hand, the English-language news media choose Fear and Trust. In particular, Fear becomes ChannelNewsAsia’s clear dominant daily emotion with 81.1%. The dominant daily emotion amplifies the difference among news media, as it can be seen from how the plots in Figure 3 and 4 differ greatly from the plots in Figure 2.

3.6. Emotion Intensity Comparison (Indonesia’s News Media)
In the previous section, it has been shown that the emotion mix and dominant daily emotion in each news media may greatly differ. In this section, it is shown that the intensity of the emotions may be
significantly different among news media, as well. In other words, even if two media have similar emotion mix or similiary dominant daily emotion, one media may be more emotionally intense than the other.

A Z-test is performed using a Python package called statsmodels [14] to compare the average daily emotion between pairs of news media with respect to a particular emotion. The result for Sadness comparison among Indonesia’s news media is shown in Table 4. The column H1 denotes the sign in the alternate hypothesis and the p-value that is less than 0.05 is considered as statistically significant. Table 4 indicates that kompascom is the most intense in Sadness, followed by republikaonline, CNNIndonesia, and jakpost. Interestingly, despite CNNIndonesia having the highest portion of Sadness in its dominant daily emotion, its intensity is significantly less than two other media.

The result in Table 4 is confirmed by the boxplot visualization of month-by-month daily emotion of Sadness of each news media in Figure 5. The Y-axis indicates the Sadness emotion score, while the labels on X-axis indicate the media name and the month, e.g., “J2” means jakpost in month 2. Meanwhile, “C” represents CNNIndonesia, “R” represents republikaonline, and “K” represents kompascom. There is no boxplot in J1 because the earliest tweets for jakpost was from February. It may be observed that kompascom’s scores for Sadness soar above other news media in all months.

Interestingly, kompascom is also the most intense in Joy, followed by jakpost and republikaonline which are not significantly different, and CNNIndonesia is the last. This matches the fact that the dominant daily emotion of CNNIndonesia is Sadness. For the other two emotions, i.e., Fear and Trust, kompascom is also significantly more intense than the others. Due to page limitation, the Z-test results and boxplots are not displayed here.
Table 5. Z-test for comparing the average daily emotion of Fear among English-language news media

| Media 1      | H1        | Media 2     | Z-value | p-value |
|--------------|-----------|-------------|---------|---------|
| jakpost      | <         | XHNews      | -11.97  | 2.47E-33|
| jakpost      | <         | timesofindia| -3.40   | 3.42E-04|
| jakpost      | >         | ChannelNewsAsia| 1.48    | 6.94E-02|
| XHNews       | >         | timesofindia| 6.18    | 3.15E-10|
| XHNews       | >         | ChannelNewsAsia| 18.78   | 5.22E-79|
| timesofindia | >         | ChannelNewsAsia| 5.27    | 6.69E-08|

Figure 5. Comparison of Sadness in Indonesia’s News Media

Figure 6. Comparison of Fear in English-language News Media
3.7. Emotion Intensity Comparison (English-language News Media)

Among English-language news media, Table 5 indicates that XHNews has the most intense emotion in Fear, followed by timesofindia, and finally both jakpost and ChannelNewsAsia do not differ significantly. Correspondingly, Figure 6 shows that XHNews’ scores for Fear are above others in most of the months. The “J”, “X”, “T”, “C” labels in the X-axis of Figure 6 represent jakpost, XHNews, timesofindia, and ChannelNewsAsia, respectively; followed by a number that indicate a particular month. It is worth noting here that ChannelNewsAsia’s daily emotion of Fear is relatively stable in all months. Despite its dominant daily emotion is Fear, ChannelNewsAsia delivers relatively low intensity of Fear compared to other news media. Furthermore, the intensity neither increases nor decreases significantly along time.

Interestingly, ChannelNewsAsia also becomes the news media that is the least intense in the other three emotions, i.e., Joy, Sadness, and Trust. XHNews is the most intense in both Sadness and Trust. XHNews and timesofindia do not differ in the intensity of Joy.

4. Conclusion and future work

This research uses the NRC Affect Intensity Lexicon to analyze the emotions within the news tweets from various news media. Using statistical testing and data visualization, it may be concluded that news media have different emotional communication strategies. It is shown by the differences in Emotion Mix, Dominant Daily Emotion, and Emotion Intensity.

The contribution of this research is that readers may be aware of the differences in emotion mix and intensity among news media, such that they may make an informed decision on reading the news tweets from a particular media. For the news media, the result in this research may be used to reflect their emotional communication strategy.

For the future work, to address the limitation of the subjective modifications that are proposed in this paper, a more objective and automatic approach may be used to modify a sentiment lexicon according to the context of the corpus. For example, word embedding technique, such as word2vec [15], may be useful to automatically identify that the word “positive” in the COVID-19 corpus generally has a negative meaning. Adding the data such as time of day, the number of likes and retweets; this research may be the basis of the emotion mix optimization to achieve the objective such as maximizing the number of readers. This research may also become the initial step in linking the emotions conveyed by the news media to the other issues, such as cultural background, emotion in different language, etc. In addition, analogous to the research about the changes in emotion expression in the developmental context [16], a time series analysis may be conducted to reveal the trend of COVID-19 news tweets along with the development of the pandemic.

References

[1] Lin K H-Y, Yang C and Chen H-H 2007 What emotions do news articles trigger in their readers? SIGIR 2007 Proc. Amsterdam, The Netherlands 23–27 July pp 733–4

[2] Zhu C, Zhu H, Ge Y, Chen E and Liu Q 2014 Tracking the evolution of social emotions: A time-aware topic modeling perspective 2014 IEEE Int. Conf. on Data Mining Shenzhen, China 14-17 December pp 697–706

[3] Kim H J and Cameron G T 2011 Emotions matter in crisis: The role of anger and sadness in the publics' response to crisis news framing and corporate crisis response Communication Research 38 pp 826–55

[4] Nabi R L 2003 Exploring the framing effects of emotion: Do discrete emotions differentially influence information accessibility, information seeking, and policy preference? Communication Research 30 pp 224–47

[5] Boukes M and Vliegenthart R 2017 News consumption and its unpleasant side effect J. of Media Psychology 29(3) pp 137–47

[6] Wirth W and Schramm H 2005 Media and Emotions Communication Research Trends 24(3) pp 3–43
[7] Harcup T and O’Neill D 2001 What is news? Galtung and Ruge revisited. *Journalism Studies* 2(2) pp 261–80

[8] Medhat W, Hassan A, Korashy H 2014 Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng. J.* 5(4) pp 1093–113

[9] Ravi K and Ravi V 2015 A survey on opinion mining and sentiment analysis: Tasks, approaches and applications *Knowledge-Based Systems* 89 pp 14–46

[10] Mohammad S M 2018 Word affect intensities *Proc. of the Eleventh Int. Conf. on Language Resources and Evaluation (LREC 2018)* Miyazaki, Japan 7-12 May pp 174–83

[11] Mohammad S M, Bravo-Marquez F, Salameh M and Kiritchenko S 2018 SemEval-2018 Task 1: affect in tweets *Proc. of the 12th Int. Workshop on Semantic Evaluation (SemEval-2018)* New Orleans Louisiana 5-6 June pp 1–17

[12] Plutchik R 1980 A general psychoevolutionary theory of emotion *Theories of Emotion* pp 3–33

[13] Hunter J D 2007 Matplotlib: A 2D graphics environment *Computing in Sci. & Eng.* 9(3) pp 90–5

[14] Seabold S and Perktold J 2010 Statsmodels: econometric and statistical modeling with Python *Proc. of the 9th Python in Sci. Conf.* Austin, USA 28 June – 3 July pp 92–6

[15] Řehůřek R and Sojka P 2010 Software framework for topic modelling with large corpora *LREC 2010 Workshop on New Challenges for NLP Frameworks* Valletta, Malta 17-23 May pp 45–50

[16] Hipson W E 2019 Using sentiment analysis to detect affect in children’s and adolescents’ poetry *Int. J. of Behavioral Development* pp 1-8