Pakistan Journal of Engineering and Technology
ISSN (Print): 2664-2042; ISSN (Online): 2664-2050
Received: 2023-05-27
Accepted: 2024-01-11
https://doi.org/10.51846/vol6iss4pp7-12

PAKJET
Pakistan Journal of Engineering & Technology

Article

Development and Evaluation of Gold Standard Dataset for Sentiment Analysis of Tweets

Saad Ahmed [1,*], Saman Hina [2], Raheela Asif [2], Sana Ahmed [3], and Munad Ahmed [3]

[1] Department of Computer Science, Iqra University, Karachi, Pakistan
[2] Department of Computer Science and Information Technology, NED University of Engineering & Technology, Karachi, Pakistan
[3] Research Department, MSM360pk, Karachi, Pakistan
* Correspondence: Saad Ahmed (saadahmed@iqra.edu.pk)

Abstract: Pre-labeled data is typically required for supervised machine learning. A limited number of object classes in the majority of open access and pre-annotated datasets make them unsuitable for certain tasks, even though they are readily available for training machine learning algorithms. For custom models, previously available pre-annotated data is typically insufficient, so gathering and preparing training data is necessary for the majority of real-world applications. The quantity and quality of annotations clearly trade-off with one another. Either more annotated data can be produced or better data quality can be guaranteed by allocating time and resources. Development of the gold standard by annotating textual information is an essential part of the "Text Analytics" domain in the field of "Natural Language Processing-NLP". In "Text Analytics", annotation can be done by adopting a manual, semi-automatic or automatic approach. In the case of the manual approach, annotators often work with partial parts of the corpus, and the results are generalized by automated text classification which may affect the final classification results. Annotations, reliability, and suitability of assigned labels are particularly important in the NLP applications related to opinion mining or sentiment analysis. In this research study, we have evaluated the significance of the annotation process on a novel dataset that contained multiple languages (English, Roman Urdu), a free text dataset that was extracted from Twitter. This unique dataset contained multiple languages which makes this annotation process essential for researching this data. Using this multi-language dataset, we examine the inter-annotator agreement in multiclass and multi-label sentiment annotation. To scrutinize the reliability of this research work, several annotation agreement metrics, statistical analysis, and Machine Learning methods have been considered to evaluate the accuracy of resulting annotations. It was observed that the annotation process is significant and a complex step that is essential for the proper implementation of Natural Language processing tasks for text analytics in machine learning. During this research, different gaps were identified and resolved which can impact the overall reliability of the annotation process which are reported in this paper. We conclude that while inaccurate annotations worsen the results, the impact is minimal, at least when using text data. The advantages of the larger annotated data set (obtained by employing subpar auto-annotation techniques) surpass the degradation resulting from the use of annotated data.

Keywords: Annotation, Sentiment Analysis, Machine Learning, Gold Standard Datasets

1. Introduction

1.1. Background and Motivation

Sentiment analysis is a type of narrow semantic analysis of texts. The goal is to extract opinions, feelings, or attitudes toward different entities [1], [2]. For instance, one might be interested in consumers' views about products or services provided by a company, investors' anticipations about stocks, or it can be a voter's outlook toward political parties. From the initial research that has been reported in the 2000s [3–6], sentiment analysis and opinion mining have gained significant importance as the "World Wide Web" and "Social Media" platforms produced enormous data growth during this period. Diverse forms and types of textual information became easily accessible. For instance, news, blogs, product reviews, Instagram, Facebook comments, and Twitter tweets. During the last decade, different approaches to sentiment analysis have been developed and are still very popular in the research conducted by different data scientists.

1.2. Literature Review

There are two predominant approaches to sentiment analysis in this era of big data.

1) Lexicon-based [7]
2) Machine learning [8]

The first approach uses a set of words that carry the sentiments of the opinion holder and the sentiment in the text is computed from those sets of words, which are identified in the text during the process of analysis. The second approach uses a model created from a large set of sentiment-labeled texts; then it is applied to the stream of unlabeled text documents. These models have a function that links the features obtained from the text into sentiment labels, which have distinct values of positive, negative, and neutral. Both approaches require the substantial involvement of human annotators. Depending on the nature of research work, human annotators have to label their view of the opinion or sentiment expressed either in the form of individual words or

can be described in short texts. This process of annotation labeling is based on annotation guidelines that are required specific to the research task. In the case of "Sentiment Analysis" of product reviews, annotators may be asked to annotate individual reviews with two classes/labels (e.g., Positive or Negative). Reference [9] describes a lexicon-based approach example that involves a massive human sentiment labeling of words. Five million human sentiment assessments of 10,000 common words—each in ten languages—were gathered, and each word was labeled fifty times. SentiWordNet [10] is another well-known sentiment lexicon that was limited to English and was created for over 100,000 words, semiautomatically. Authors in [11], [12] have expressed that there is always a need for an accurately annotated dataset, training a Machine learning classifier with such a dataset is vital for the success of the model [13]. Opinion mining and Sentiment analysis has gained widespread popularity and a vast raging applicability hence numerous approaches are addressing the task [14], [15]. The researchers in [16] pointed out the important role of high-quality annotations for training data and the effects of the inter-annotator agreement on the performance of the ML classifier. Also examined in [17] was the topic of agreement across annotators when it came to multi-class sentiment annotations. This researcher worked with 3255 documents in German language which were collected from different social media networks. On average, each document had 50 words, so restricting the subjects and ideas that each message can express. The following labels were applied to the texts by six annotators: negative, neutral, positive, no sentiment, irrelevant, and undecided. The inter-annotator agreements between every pair of the six annotators were estimated using Cohen's kappa. The kappa values were 0.480 for the worst and 0.747 for the best. To choose the final labels for a message, an algorithm based on majority vote was employed. When applied to a sentence-based annotation of a Modern Standard Arabic newswire sentiment dataset, a research study in [18] proposed a multi-class sentiment annotation schema based on the news domain and achieved 88.06% agreement between two annotators with a Kappa value of 0.823. Manual annotations and automated annotations were compared [19]. The writers employed a corpus of 1787 sentences gathered from transcripts of hearings conducted by a Senate Committee in the United States. An average of three labels were obtained for each sentence by the four annotators who could annotate it into ten possible labels; the obtained agreement was 0.30. The sentences were then classified using a k-NN algorithm, which produced an F-measure of 0.4. Subsequently, the automated classification method that determines F-measures against the final annotation also referred to as the "ground truth" (gold standard), was applied to the evaluation of human annotations. 0.70 was the ideal F-measure for human annotation. Sixty non-healthcare related Master's students were recruited to manually annotate 600 sentences collected from the English-language Spine-health forum by another team of researchers [20]. A total of 6 basic Emotions were used by annotators [21]: anger, disgust, fear, joy, sadness, and surprise. Two annotators annotated each sentence. Kappa, the coefficient of inter-annotation, was 0.26. Two medical professionals annotated 150 sentences from the same corpus. A moderate 0.46 agreement was found between annotators who were health professionals and nonprofessionals. Despite this, the authors continued with their machine-learning experiments; their highest F-score was 0.65. Annotated were 150 topics from 115 documents (from the TREC 2008 Blog Track) [22]. For every topic, an average of 3.6 annotators were engaged. Krippendorf alpha was used to assess the annotators' agreement [23]. The measure [24] accepts non-annotated examples and, in contrast to Cohen and Fleiss kappa, can evaluate agreement among a variable number of annotators; the authors considered this kind of agreement to be moderate. Reference [22] stressed the need for further studies of different levels of annotations, i.e., document, paragraph, and sentence levels. The significance of human-machine collaboration for sentiment analysis and determining the degree of agreement between several human and machine annotators was highlighted in a more recent study called "human-in-the-loop." References [25] and [26] analyzed the performance and agreement between off-the-shelf sentiment analysis tools and reported the sentiment measurement done by an average of 5.63 coders proved satisfactory reliability of Krippendorff's alpha = .80, where the assessment was made on a sample of 148 randomly selected news items analyzing the sentiment of newspaper and website headlines, manually annotated by a team of 22 student coders who were initially trained. The research in [27] focused on more expressive annotations by conducting a two-phase annotation arrangement and showed that perceived emotions can be different from expressed emotions in an event-focused corpus, in turn affecting the performance of the classifier. Another team of researchers manually annotated 7,000 tweets for each of the seven emotions. They then selected 14 topics that they thought would provoke emotional tweets, gathering hashtags to make it easier to find tweets about these subjects. The annotators achieved alpha = 0.67 inter-annotator agreement on 500 tweets after multiple iterations [28]. In this research, a set of Twitter tweets collected through Twitter's "Application Programming InterfaceAPI" were analyzed. These tweets were retrieved as a mixture of multiple languages (English and Roman Urdu language). These tweets were then manually annotated by three Human Annotators. The annotated tweets were used to train sentiment classifiers in the form of training data. This type of dataset was also used in our research which is described in [29]. A synopsis of the state-of-the-art Twitter sentiment analysis is reported in [30]. Another study on lexicon-based machine learning techniques and their integration is provided in [31]. In their study results, authors [32] emphasized how crucial it is to validate automatic text analysis techniques before using them. Authors in [33] have highlighted that Social reviews are more complex and it is difficult to extract aspects from them without using properly annotated training datasets. The number and caliber of the labeled tweets have been our main concerns. The consensus

amongst human annotators is used to assess the quality of the labeled tweets.

2. The Dataset

To build our dataset, we collected tweets from twitter.com using the Twitter API; tweets collected were related to the telecommunication domain. These tweets were collected from the official handles of major telecommunication companies in Pakistan. The total number of collected tweets was '4123', then the Retweets and tweets were filtered out that only contained URLs or which were without any message. After this cleaning process, '2703' tweets were retained in our dataset, statistics of the dataset are shown in Table 1. This final dataset is used in this research work.

Table 1. Statistics of dataset.

Dataset	Total Collected Tweets	Total Removed Tweets	Total Tweets Retained
Twitter Reviews	4123	1405	2718

3. Materials and Methods

We used three sentiment labels for this research work: Positive, Negative, and Neutral. As one tweet may convey more than one sentiment which can create confusion, we decided to label one sentiment for each tweet which covers the overall sentiment polarity of the tweet. Each document was annotated by three annotators separately. The annotators who participated in this research work were three undergraduate students of the "Computer Science" department of the university. They were selected for this research based on their language proficiency in both English and Urdu languages. The next section includes the process of annotation based on flexible guidelines provided to annotators.

3.1. Initial Text Annotation

In the annotation guidelines that were provided to the annotators, three sentiment labels/categories (Neutral, Negative, and Positive) were provided in the dataset as default labels. It was then suggested to annotators that they could designate as many sentiments as they thought appropriate for each message. Even while annotators could use the "multi-label" feature to give many labels to a post, they often just used one predetermined label: approximately 91% of the assigned labels were from the pre-defined set. New labels suggested by the annotators are sad, disgusted, satisfied, happy, and surprised. For current research work, we only consider the three default labels, but we intentionally collected the additional labels suggested by annotators and we plan to apply these new suggested labels in our future research work of upgrading our earlier work "advanced framework based on aspect-based sentiment analysis" [34].

3.2. Final Text Annotation

We used the following statistics to select a label for every tweet based on all the labels annotators had assigned:

- The same label for a tweet was assigned by all three annotators. This happened in the case of 2136 tweets.
- Non-matching labels were also assigned too this happened for 472 tweets. The label given by two of the three annotators was selected as the final in such cases.
- All three annotators were assigned two labels for an identical tweet; the second label was optional for annotators. This happened for 64 tweets.
- All three annotators assign different labels to a tweet. This happened for 46 tweets. These tweets were given the final label after the agreement between the three annotators.

With an average of 1.07 labels per tweet and 80% of tweets annotated with just one label, annotator-A often added fewer labels. With an average of 1.25 labels per tweet and 65% of tweets annotated with just one label, Annotator-B and Annotator-C added more labels.

3.3. Assessment for Inter-Annotator Agreement

For inter-annotators agreement, several metrics were used for the evaluation in this research work. The Percent of agreement is the most straightforward metric that provides a basic approximation of overall agreement between annotators. These are shown in Table 2.

Table 2. Inter-annotator percentage of agreement.

Dataset	Positive	Negative	Neutral
ANO1 Vs ANO2	87.4%	78.6%	77.1%
ANO2 Vs ANO3	84.2%	76.2%	75.3%
ANO1 Vs ANO3	89.7%	79.0%	77.7%

A total of three labels were used, with inter-annotator agreement calculated for each, because every tweet includes several labels. The conditions are the same for the three annotators. Table 2 displays the averages and the pair-wise agreement for each pair of annotators. There is strong agreement, as demonstrated by the average agreement of 89.7%. There is nearly constant agreement between the three pairs of annotators. For this reason, it cannot be assumed that one annotator is better than others. While comparing per labels agreement we found that annotators mostly agreed on Positive Sentiment Polarity and lesser agreement was made on Negative and Neutral. This may be due to the nature of the text present in tweets which are limited to only 140 characters. Krippendrof's alpha, Fleiss Kappa (K), and Cohen's Kappa (k) are used to assess inter-annotator agreement. Since Cohen's Kappa and Fleiss Kappa are measures above chance agreement, they are chance-corrected coefficients. These coefficients' formula is

$$\alpha = \frac{(\mathrm{Ao} - \mathrm{Ae})}{(1 - \mathrm{Ae})} \quad (1)$$

When the annotators choose the labels at random, Ao is the observed agreement and Ae is the expected agreement.

3.4. Cohen's Kappa

The agreement between two raters who place N items into C mutually exclusive categories is measured by Cohen's kappa. The definition of k is

$$\kappa = \frac{(\mathrm{Po} - \mathrm{Pe})}{(1 - \mathrm{Pe})} = 1 - \frac{1 - \mathrm{Po}}{1 - \mathrm{Pe}} \quad (2)$$

The anticipated agreement in Cohen's K Ae is computed under the supposition that the prior distributions, distinct to each annotator and observed from their actual distribution, govern the random assignment of categories to the item. Cohen k was computed for every pair of our annotators, and it is only applied to two of them. Average Cohen's k = 0.45 indicated moderate agreement. Table 3 displays agreements between annotators.

Table 3. Cohen Kappa agreement between annotators.

Dataset	Positive	Negative	Neutral
ANO1 Vs ANO2	0.44	0.43	0.50
ANO2 Vs ANO3	0.45	0.40	0.55
ANO1 Vs ANO3	0.53	0.42	0.48

3.5. Fleiss' Kappa

When categorizing items or assigning ratings to multiple items, the Fleiss' kappa statistic—named after Joseph L. Fleiss—is utilized to evaluate the consistency of the rating given by a fixed group of raters.

$$\kappa = \frac{(\bar{\mathrm{P}}\mathrm{o} - \bar{\mathrm{P}}\mathrm{e})}{(1 - \bar{\mathrm{P}}\mathrm{e})} \quad (3)$$

The degree of agreement that can be achieved above chance is given by the factor $1 - \bar{\mathrm{P}}\mathrm{e}$, while the degree of agreement that has been achieved above chance is given by $\bar{\mathrm{P}}\mathrm{o} - \bar{\mathrm{P}}\mathrm{e}$. In the event that all raters concur, then $ka\ p\ pa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $ka\ p\ pa \leq 0$. This anticipated agreement, which is a generalization for more than two annotators, is computed under the presumption that the distribution of items among categories in the real world governs any annotator's random assignment of categories to items. The results are displayed in Table 4.

Table 4. Fleiss Kappa.

Dataset	Positive	Negative	Neutral
Fleiss K	0.46	0.39	0.47
Observed	0.71	0.82	0.88
Expected	0.52	0.54	0.65

3.6. Krippendorff's Alpha

It is an agreement coefficient that is predicated on the idea that expected agreement is determined by examining the judgment distribution as a whole, regardless of the annotator who generated the judgments. Alpha is given by:

$$\alpha = 1 - \frac{\mathrm{Do}}{\mathrm{De}} \quad (4)$$

where Do is the disagreement observed and De is the disagreement expected by chance. It also applies to multiple annotators and also allows missing values. Alpha is calculated for each label, and then it is averaged, as shown in Table 5.

Table 5. Krippendorf's alpha.

	Positive	Negative	Neutral	Average
Alpha	0.44	0.46	0.46	0.45

In all computed measures, the average agreement is 0.45, which is regarded as moderate. In this current study, there was no visible difference between the results but we can infer from our pair-wise metrics, which annotator is best among all annotators who worked on this corpus. Annotator Agreement per label was calculated which shows us the label identification difficulty, The best results of the agreement are for labels that are easier to detect while the worst results are an indication of labels that are hard to detect by annotators. The Gold Standard Dataset Preparation: After going through all the investigations on the dataset, the differences of annotators on tweet labels were resolved by adopting a procedure presented in Fig. 1. In this process of resolving disagreements, all the tweets that were labeled differently by annotators were discussed in the presence of annotators, and their viewpoints were taken into account before assigning the final label to these tweets. In this way, the differences were resolved and a dataset with labels was finalized, which is then used to train Machine Learning Algorithms for further processing.

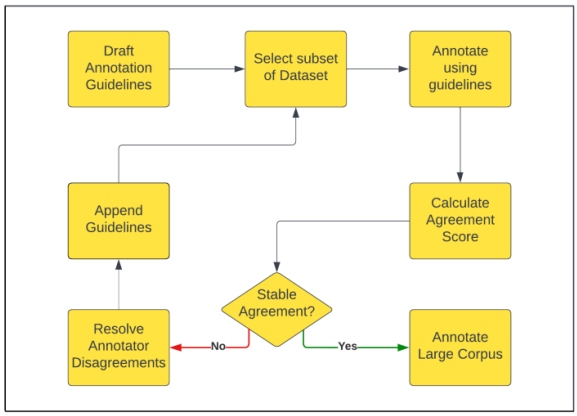

Figure 1. Annotation process flow chart.

We used different statistical methods to evaluate the performance of annotation tasks using several inter-annotator agreements (Krippendorff's Alpha, Fleiss Kappa, and Cohen's Kappa) that are known for the evaluation of the developed GOLD standard dataset. The role of human annotators is very vital and the selection of human annotators to work on a particular dataset is very tricky because it must make sure that the human annotator selected to work on a dataset must have consistent knowledge and expertise of the domain and must be trained to be reliable and competent in

their work which assures that dataset is annotated consistently and is reliable enough to be tested on an automated system.

The results of polarity with percentages of the annotated dataset are shown in Table 6. The ML model used to determine the performance of the normally labeled dataset and the gold standard dataset is depicted in Fig. 2.

Table 6. Statistics of annotated dataset.

Dataset	Total	Positive	Negative	Neutral
Twitter Reviews	2073	1735	224	745

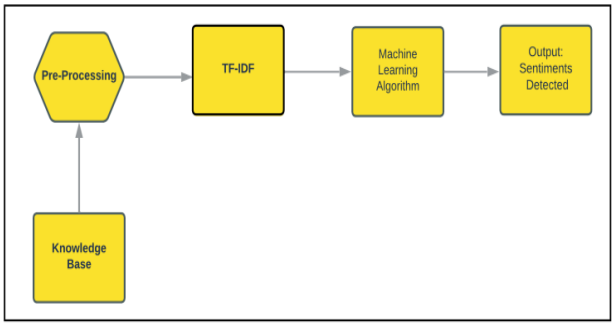

Figure 2. Machine learning model used to evaluate our annotated dataset.

After testing both datasets, it was evident that the gold standard dataset was indeed producing better results and the accuracy of the same ML model was improved which proves that the overall performance of the annotator in applying the guidelines with ample domain knowledge during the annotation of the dataset will impact the ML model positively. The results are shown in Table 7 and are also depicted in Fig. 3.

Table 7. Performance using different ml algorithms.

Algorithms	Performance using Normal Annotated Dataset		Performance using Gold Standard Dataset	
	Accuracy	F Score	Accuracy	F Score
Naive Bayes	0.801	0.791	0.811	0.824
SVM	0.691	0.551	0.721	0.674
Forest	0.791	0.668	0.812	0.765
Tree	0.251	0.211	0.314	0.234

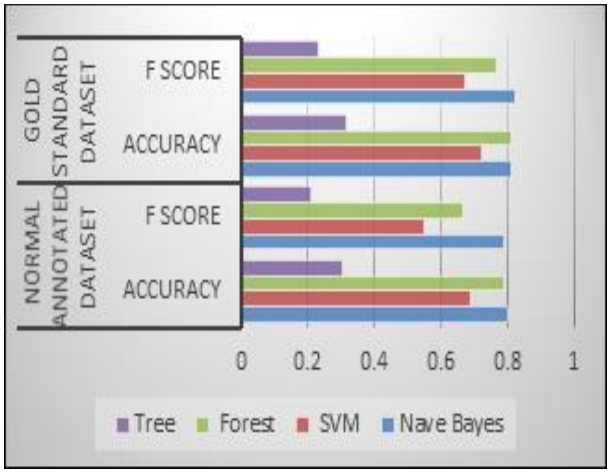

Figure 3. Machine learning model performance.

4. Conclusion

This study used a multi-language dataset to investigate multi-class sentiment annotation. The dataset was collected from the website of the online social media network Twitter. To estimate the quality of annotation, different inter-annotator agreement metrics were applied. It was observed how these metrics can be applied for the evaluation of sentiment categories and also for keeping the high quality of annotated data. This research provided new insights and encouraging results and contributed a gold standard dataset. The gold standard was finalized after adopting a procedure in which disagreements between annotators were resolved and a final label was assigned to the tweet (which previously was annotated with multiple labels by the annotators). Statistical metrics were used to evaluate this research work because this dataset is unique in the sense that it was collected from social media on a specific domain that was related to telecommunication networks.

In this research study, a novel dataset was used that contained multiple languages (English and Roman Urdu), and on this dataset, the significance of the annotation process was evaluated concerning its importance when used to train a machine learning model to analyze text written in multiple languages simultaneously. To evaluate the scalability of this research work, inter-annotator agreement in multi-class and multi-label sentiment annotation using this multi-language dataset was examined. The accuracy of resulting annotations was evaluated by considering several annotation agreement metrics and statistical analysis. It was witnessed that annotation is a significant and complex process and when the annotated dataset is used for training a machine learning model, this process of annotation becomes essential for the accurate and reliable implementation of Natural Language processing tasks for text analytics of huge unstructured data being generated on social media on daily basis. We plan to expand this research work to new datasets from other domains. We aim to investigate and identify procedures and conventions for the generalization of the annotation process for sentiment analysis.

References

[1] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, Art. no. 1, 2012.

[2] J. Zhao, K. Liu, and L. Xu, "Sentiment analysis: mining opinions, sentiments, and emotions," MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info, 2016.

[3] T. Briscoe and J. A. Carroll, "Robust accurate statistical annotation of general text.," in *LREC*, 2002.

[4] M. Delgado, M. J. Martín-Bautista, D. Sánchez, and M. Vila, "Mining text data: special features and patterns," Springer, 2000, pp. 140–153.

[5] A. B. Brush, D. Bargeron, A. Gupta, and J. J. Cadiz, "Robust annotation positioning in digital documents," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2001, pp. 285–292.

[6] P. Ferguson *et al.*, "Exploring the use of paragraph-level annotations for sentiment analysis of financial blogs," 2009.

[7] R. M. Sallam, M. Hussein, and H. M. Mousa, "Improving collaborative filtering using lexicon-based sentiment analysis," *International Journal of Electrical and Computer Engineering*, vol. 12, Art. no. 2, 2022.

[8] G. Revathy, S. A. Alghamdi, Alahmari, Sultan M, Yonbawi, Saud R,

A. Kumar, and M. A. Haq, “Sentiment analysis using machine learning: Progress in the machine intelligence for data science,” *Sustainable Energy Technologies and Assessments*, vol. 53, p. 102557, 2022.

[9] P. S. Dodds *et al.*, “Human language reveals a universal positivity bias,” *Proceedings of the national academy of sciences*, vol. 112, Art. no. 8, 2015.

[10] A. Esuli and F. Sebastiani, “Sentiwordnet: A publicly available lexical resource for opinion mining,” in *Proceedings of the 5th international conference on language resources and evaluation*, 2006.

[11] M. Alghobiri, “A comparative analysis of classification algorithms on diverse datasets,” *Engineering, Technology & Applied Science Research*, vol. 8, Art. no. 2, 2018.

[12] Z. A. Shaikh, “Keyword detection techniques,” *Engineering, Technology & Applied Science Research*, vol. 8, Art. no. 1, 2018.

[13] K. Alhazmi, W. Alsumari, I. Seppo, L. Podkuiko, and M. Simon, “Effects of annotation quality on model performance,” in *IEEE*, 2021, pp. 063–067.

[14] R. Bakshi, N. Kaur, R. Kaur, and G. Kaur, “Opinion mining and sentiment analysis. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) 2016 Mar 16 (pp. 452-455),” IEEE.

[15] B. Liu and L. Zhang, “A survey of opinion mining and sentiment analysis,” Springer, 2012, pp. 415–463.

[16] I. Mozetič, M. Grčar, and J. Smailović, “Multilingual Twitter sentiment classification: The role of human annotators,” *PloS one*, vol. 11, Art. no. 5, 2016.

[17] J. Yang, Y. Zhang, L. Li, and X. Li, “YEDDA: A lightweight collaborative text span annotation tool,” 2017.

[18] M. Abdul-Mageed and M. Diab, “Subjectivity and sentiment annotation of modern standard arabic newswire,” Proceedings of the 5th linguistic annotation workshop, 2011, pp. 110–118.

[19] P. Wang, E. Ishita, and D. W. Oard, “Kenneth r. Fleischmann.”

[20] S. Melzi, A. Abdaoui, J. Azé, S. Bringay, P. Poncelet, and F. Galtier, “Patient’s rationale: Patient Knowledge retrieval from health forums,” in *eTELEMED: eHealth, Telemedicine, and Social Medicine*, 2014.

[21] P. Ekman, “Are there basic emotions?,” American Psychological Association, 1992.

[22] A. Bermingham and A. F. Smeaton, “A study of inter-annotator agreement for opinion retrieval,” in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 784–785.

[23] A. F. Hayes and K. Krippendorff, “Answering the call for a standard reliability measure for coding data,” *Communication methods and measures*, vol. 1, Art. no. 1, 2007.

[24] R. Falotico and P. Quatto, “Fleiss’ kappa statistic without paradoxes,” *Quality & Quantity*, vol. 49, Art. no. 2, 2015.

[25] Yeruva, Vijaya Kumari, M. Chandrashekar, Y. Lee, J. Rydberg-Cox, V. Blanton, and N. A. Oyler, “Interpretation of sentiment analysis with human-in-the-loop,” in *IEEE*, 2020, pp. 3099–3108.

[26] M. Boukes, Van, T. Araujo, and R. Vliegenthart, “What’s the tone? Easy doesn’t do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools,” *Communication Methods and Measures*, vol. 14, Art. no. 2, 2020.

[27] E. Troiano, S. Padó, and R. Klinger, “Crowdsourcing and validating event-focused emotion corpora for German and English,” 2019.

[28] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. Harabagiu, “Empatweet: Annotating and detecting emotions on twitter,” in *Lrec*, 2012, pp. 3806–3813.

[29] A. Saad, S. Hina, and R. Asif, “Detection of sentiment polarity of unstructured multi-language text from social media,” *International Journal of Advanced Computer Science and Applications*, vol. 9, Art. no. 7, 2018.

[30] E. Martínez-Cámara, M.-V. M. Teresa, U.-L. L. Alfonso, and M.-R. A. Rturo, “Sentiment analysis in twitter,” *Natural language engineering*, vol. 20, Art. no. 1, 2014.

[31] O. Kolchyna, T. T. Souza, P. Treleaven, and T. Aste, “Twitter sentiment analysis: Lexicon method, machine learning method and their combination,” 2015.

[32] V. Atteveldt, Van, and M. Boukes, “The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms,” *Communication Methods and Measures*, vol. 15, Art. no. 2, 2021.

[33] J. Mir, A. Mahmood, and S. Khatoon, “Aspect based classification model for social,” *Engineering, Technology & Applied Science Research*, vol. 7, Art. no. 6, 2017.

[34] S. Ahmed, S. Haman, E. Atwell, and F. Ahmed, “Aspect based sentiment analysis framework using data from social media network,” *IJCSNS Int. J. Comput. Sci. Netw. Secur*, vol. 17, pp. 100–105, 2017.