FOODIE: A Data-centric Sifting Framework for Social Media Analytics

Dolly Agarwal*
Ai Palette Pte Ltd
Singapore
dolly@aipalette.com

Abdullah Al Imran†
Ai Palette Pte Ltd
Singapore
imran@aipalette.com

Kasun S Perera‡
Ai Palette Pte Ltd
Singapore
kasun@aipalette.com

*ORCID: 0000-0001-8004-8650
†ORCID: 0000-0002-3781-8178
‡ORCID: 0009-0005-3076-1059

***Abstract*—There has been a great deal of research conducted in the past on utilizing social media analytics to derive consumer insights and understand their behaviors. However, when such studies are applied to real-world data in an industrial use-case, the results are often found to be incorrect and erroneous. This is a major barrier for companies that provide social media analytics-based solutions to customer-centric industries such as Food and Beverage (FnB). One of the key causes of this barrier is the failure to appropriately process and curate raw social data prior to analytics. In this study, we discuss the challenges we encountered when dealing with social data throughout our industrial experience and propose a standard solution - FOODIE. This is a framework specifically designed for the FnB industry to process social conversation data accurately and in a standard manner prior to perform various downstream tasks on it. The three stages of this paradigm are preparation, sifting, and evaluation. Through this framework, we have reduced the data to error ratio from 8.76% to 0.01% which is quite significant given the volume of the data. While this framework is designed for the FnB space, it can be customized to suit the needs of various other industries.**

***Index Terms*— Data-centric AI, Social Media Analytics, Data Engineering, Sifting Framework, Natural Language Processing (NLP), Classification, Transfer Learning, Sentence Transformers, Clustering, Machine Learning (ML), Triplet Loss, Topic Modeling**

I. Introduction

Social media analytics in food and nutrition has recently gained attention [24]. People love to document, photograph, and share their eating habits, emotions and advices about food [3]. This serves as a crucial data source to the FnB companies for monitoring food trends, opinions, eating behaviors and market potential. Numerous studies have also been conducted on the use of algorithms to extract additional insights from social media data [7], [14], [17], [23]. However, when these methods are applied in real-world applications by businesses, they fail to deliver valid results. One common stumbling block that hampers these efforts is data quality and reliability issues. Thus, ensuring the accuracy, and trustworthiness of data is currently a major problem for business. Despite the importance of assessing and governing data quality for social media analytics, there aren't many studies that specifically address this issue [5], [20]. Moreover, the ideal way to assess the impact of poor data quality is to employ some downstream tasks, but in order to do that, we typically need a test set made up of test samples and their ground truth labels. Standard use cases meet this criteria, however many real-world instances contain test data that is not labeled, making it impossible to apply typical model evaluation techniques.

In this work, we present our industrial experience, the challenges we faced and solutions we implemented while working with social conversations to prepare it for downstream tasks like FnB trend forecasting, and innovation of potential new products. We extract social media posts from Twitter, Instagram and Facebook using food-related keywords and hashtags which is the only available way to extract social media data from such platforms. These keywords/hashtags are listed by our domain experts and research analysts. Some examples of hashtags used to extract data from USA are 'bostonfoodies', 'newyorkfoodguide', 'miamifood'. But, this method of data extraction comes with a trade-off between high precision which means employing targeted keywords to get fewer, but more accurate results; and recall which utilizes broad terms to produce a lot of irrelevant results. In reality we have experienced that it is cheap to collect the data, but expensive to annotate and sift them [7]. We examine these problems in-depth in this study and offer ML-based and adaptive rule-based solutions. However, social interactions are challenging to interpret due to the significant use of slang, typos, concatenated words, virality, and various forms of promotional content from businesses. We look into these issues as well and offer some heuristics-based solutions.

Finally, we devise an evaluation strategy to approximate precision-recall; an ideal framework should identify relevant data while returning as close to zero false positives as possible. It is worth noting that spot checking for false positives is simple and part of quality assurance process in many businesses, but there are not many ways to systematically and statistically assess false negative rates. Hence, we propose two metrics: data-to-error ratio and relevancy to approximate data quality without much annotated data.

DOI reference number: 10.18293/SEKE23-175

In summary, this paper makes the following contributions:

- We design an end-to-end framework: FOODIE to prepare social media data for downstream analytical tasks. We show how we can use heuristics and ML based approaches in the denoising and sifting process.
- We propose a novel method to evaluate social data quality quantitatively when it is expensive to get annotations and hence not straightforward to measure metrics like Precision-Recall on a downstream task.

The rest of the paper is organized as follows: Section II details the challenges addressed in this work. In Section III, we present our proposed framework, details of which is described from Section IV to Section V. Section VI explains the validation and evaluation aspect of our framework, and finally, in Section VII we conclude.

II. Challenges Addressed

Empirical investigation done in [6] reveals that a firm's intention for big data analytics can be positively affected by its competence in maintaining the quality of data. Good quality data provides better leads, better understanding of consumers and trends. Some of the challenges with social data have been looked at in [9], [14], [17], [22]. There also has been a few attempts to provide a quality framework for social media data as in [5], [20]. To the best of our knowledge, this is the first body of work that focuses on data quality challenges of social media analytics for FnB companies and provides a framework to combat it. We detail the challenges as below:

Homographs: Since the data is extracted using a keyword based approach, difficulty arises when the phrases containing homographs are encountered [2]. All posts with words that are written/spelled the same but have different meanings are considered while collecting. Some of these posts may not be food-related at all. For example - 'Nut' can refer to both dry fruit and a type of hardware tool.

Context variation: Food-related items can be used in other domains especially in skin-care. For example: Green Tea is consumed as a beverage, it can also be the ingredient of a skincare product used by customers.

Business/Promotional posts: Many businesses use social media for advertising, promotional giveaways etc. The redundant and invalid information, such as advertising posts are not the Voice of the customer (VoC) and may alter the analytics [8], [23]. Hence these posts should be identified and handled carefully.

Viral/Influencers posts: Some posts have more engagement (likes, comments, shares) than others as they may be posted by influencers or popular figures on the internet. Such posts are important to handle appropriately else they can skew trend analytics [13].

Irrelevant Hashtags: People use trending hashtags to try to reach as many users as possible, even when their content has nothing to do with the hashtag in question i.e hashtags are often misused in an effort to get more engagement. This results in a lot of False Positives in the database when we extract social posts using the hashtag based approach. For example: 'New music coming soon to all streaming platforms. #Music #NewArtist #Foodie #HealthyFood'.

Apart from these data challenges, another common roadblock is data quality assessment and impact evaluation. Experimenting on some downstream tasks is the optimum way to evaluate the effects of bad data quality, but in order to do so, we normally need annotated data which are expensive to curate. Since real-world scenarios involve data that is not labeled, it makes it difficult to employ conventional methods for model evaluation. Hence, we worked on a new evaluation metric to approximate the data quality (Section VI).

III. Proposed Framework

Figure 1 shows our framework FOODIE that is deployed in production. The overall framework is divided into 3 phases as listed below:

Data Preparation: This is the first phase that helps to remove and fix noise unique to social media data. Detail about how we do it is discussed in section IV.

Sifting: We filter out noisy and irrelevant data samples in this phase and bucket the relevant samples into food categories. We further remove duplicates to prevent over-indexing of a few samples. These steps are elaborated in section V.

Validation: Finally, we validate the data quality returned by the framework. We discuss this phase in detail in Section VI.

IV. Data Preparation Pipeline

Social conversations can have plenty of anomalies like typos, repeated characters in a word, spelling errors, concatenated words, short document lengths compared to the long text and normal documents. Inspired by the SMDCM model [17], we divided our cleaning process into two parts: Noise Removal and OOV handling. The aim of these data cleansing stage is to eliminate noise and transform the user posts into a formal standardized English language.

Noise Removal: We used regex to transform all user mention to $\langle username \rangle$, urls to $\langle url \rangle$, phone number to $\langle phoneno \rangle$. We removed all symbols, punctuations and transformed emojis/emoticons to its word representation. We also separated the text into text_only and hashtags of a post into two separate fields. If a hashtag appears within a text, we consider it as part of the text_only. For example: if a post is '#chocolate #coffee #cake made for hubbyś birthday #foodie #instafood #instafoodie #streetfood' then text_only field will have 'chocolate coffee cake for hubby s birthday' and hashtags field will have 'foodie, instafood instafoodie,

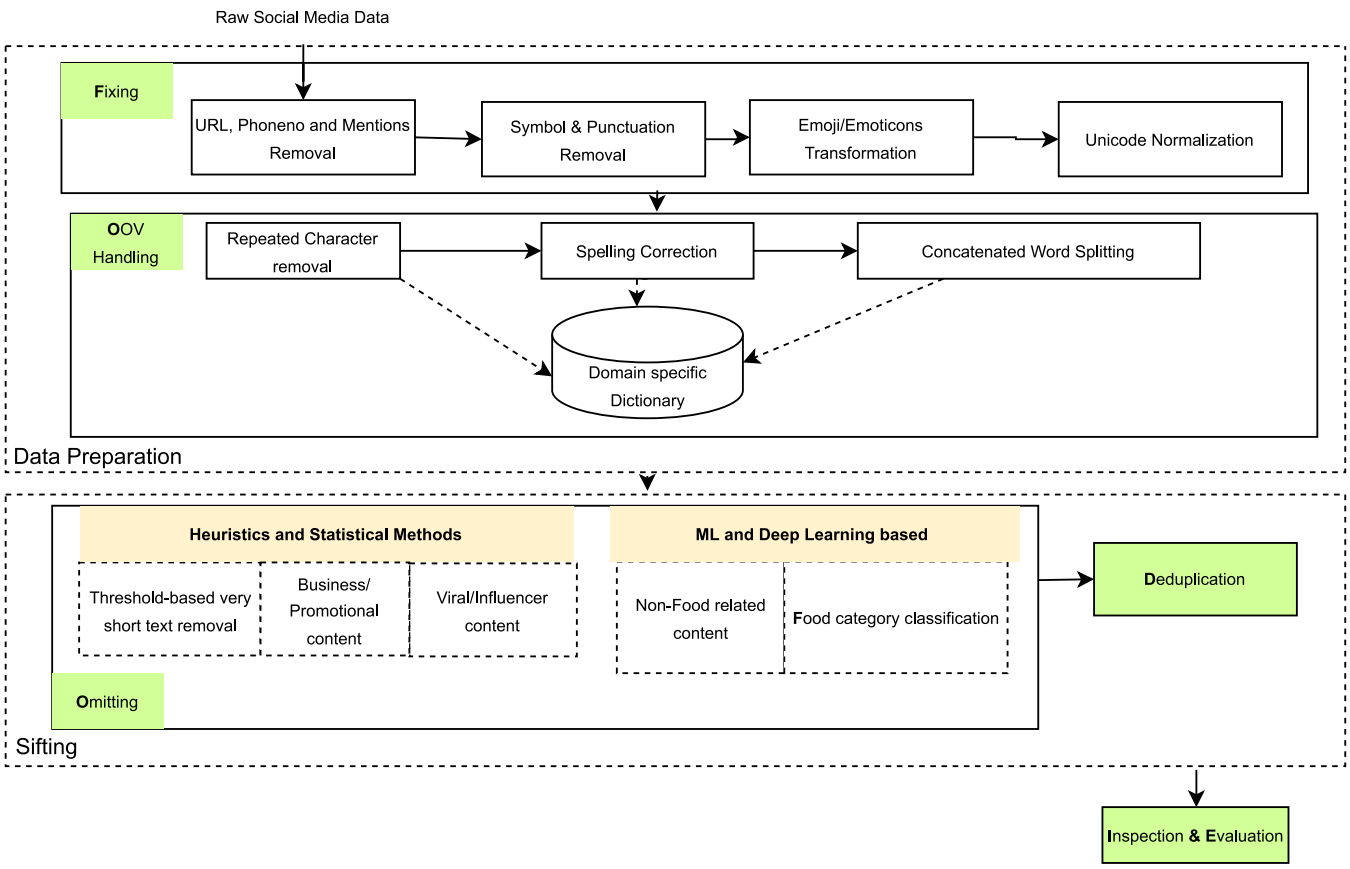

Fig. 1. FOODIE: Our Proposed Framework
Block diagram of the proposed framework

streetfood'

Out-of-Vocabulary (OOV) Handling: Words might be OOV in social media data due to reasons like, typo or use of elongated words to express feelings or concatenated words to tackle word limitation of a social platform. We first transform elongated words into their original word by eliminating repeated letters. Then we use Peter Norvig's edit distance based algorithm [1], to fix the spelling mistakes. Lastly, we used a trillion-word corpus [10] to split the concatenated words into their individual parts. For example, consider a post 'I looove Baklava. The layers of flaky filo pastyr, pistachio filling and syrup has myheart'. In this process, 'looove' is transformed to 'love', 'pastyr' gets fixed to 'pastry' and then 'myheart' splits into 'my heart'.

V. Sifting Pipeline

The purpose of this second phase is to filter out irrelevant data points and to mark those points which are true VoC related to FnB. We segregate this pipeline into two parts as described in section V-A and V-B depending on the type of approach used.

A. Heuristics and Statistics based sifting

1) Short-text removal: We considered only the text_only (without hashtags) data to infer the lower bound of text length to consider a post for analysis. Empirically we found that if a post contains ≤ 15 characters, we do not get much information about any trend that can be used for analytics, hence we filtered out all the posts with less than 15 characters.

2) Flagging promotional content: Next, we identified promotional posts using heuristics like most frequent keywords used by businesses. Promotional content have some characteristics like they usually have a url, phone number, or contain one of the keywords like 'order now', 'Call for orders', 'DM to know more'. According to frequency count, a few of the top keywords are 'buy now' (43,698), 'order now' (8,829), 'Link in bio' (7,849), 'follow us' (4,206). We also did a sanity check by looking at quotes randomly for each keyword to confirm our hypothesis. Additionally, we mark the usernames of these flagged posts and flag all other posts by those users as business posts. We flag these posts instead of removing them because they will be a useful source to analyze what local businesses or different brands are promoting in FnB.

3) Flagging viral content: Lastly, we use Outlier Detection to identify viral/influencer content. These are the posts that

```
1 top_n_words[6][:10]

[('brunch', 0.057532912565673076),
 ('breakfast', 0.045371993458157386),
 ('eggs', 0.03787316041547333),
 ('toast', 0.03386468104328588),
 ('egg', 0.028933962982201 68),
 ('bacon', 0.0203922304839994),
 ('french', 0.019938907858183256),
 ('benedict', 0.01897517788823989),
 ('pancakes', 0.017206545997296187),
 ('waffles', 0.014430150212146425)]
```

```
1 top_n_words[0][:10]

[('alinea', 0.4651009229326586),
 ('molecular', 0.2596036742097027),
 ('gastronomy', 0.2555971302221256),
 ('lifetime', 0.25026665048033997),
 ('unforgettable', 0.2427377099248482),
 ('introduction', 0.23945598488604014),
 ('starred', 0.2364289229970134),
 ('deserves', 0.23528091270942597),
 ('destination', 0.22623380330001566),
 ('michelin', 0.20172941823711418)]
```

Fig. 2. Examples of top 10 words in clusters

receive more engagements than others but may turn out to be fad. Yet, these posts can bias the overall analytics if not handled well. To spot such posts, we transformed each engagement value (likes, shares and comments) in the range 1-100. Then, we calculate overall engagement score, ES using equation 1

$$ES = \alpha * comments + \beta * shares + (1 - \alpha - \beta) * likes \quad (1)$$

where α = 0.3, β = 0.5, for Facebook and Twitter, while α = 0.4, β = 0 for Instagram. We tuned these α and β values empirically. These values account for the fact that it is easy to get likes in social platforms as compared to comments and shares. Then we marked all points with Z-score ¿ 3 or ¡ -3 to be an outlier as it indicates that the data point is 3 standard deviation away from the mean. Similar to what we did for business posts, we flag these outliers instead of directly removing them because they are still information about a food trend that we do not want to lose.

B. Machine Learning based sifting

1) Semi-supervised approach to text data labeling: To prepare the labeled data for Food/Non-Food text classification, we used Sentence Transformers [18] with model 'nli-roberta-base-v2' for feature extraction and UMAP [16] for dimensionality reduction. After reducing the dimension, we use HDBSCAN [4] to cluster the posts. The decision is based on the knowledge that Sentence Transformers performs well for tasks like semantic search and HDBSCAN can find clusters of varying densities (unlike DBSCAN), and be more robust to parameter selection. Then, we look at the top 10 words in each class based on frequency of occurrences and identify which classes are non-food. Example of two such clusters are shown in Figure 2

2) Multi-modal approach to food/non-food classification: We performed multiple experiments of FnB-related post recognition on the curated dataset using visual information, textual information and the fusion of both. We employ RoBERTa base [15] and ResNet-18 [11] for the two modalities: text and image classification respectively. To successfully apply ResNet-18 architecture to our task, we use transfer learning using the Food-5K [21] image dataset. It is a binary classifier to predict if the image in the post contains food or not. The choice of ResNet-18 comes from the willingness to reach a compromise between performances and achievable accuracy. We used semi-supervised labeled dataset (sec V-B1) to train and test a text classifier. We fine-tune a robust Transformer model RoBERTa-base using the text_only of labeled data to predict if the post is related to FnB or not. Apart from the text length concerns, we do-not consider hashtags in this approach because we found that people/businesses use trending hashtags in their posts just to gain more engagement. These trending hashtags although food-related may not be relevant to the content they are posting. Finally, we investigate various methods for performing multi-modal fusion, analyzing their trade-offs in terms of classification accuracy and computational efficiency. We adopted a late fusion approach, where the final score FS is a linear combination of the scores provided by both image and text classification systems (Equation 2). This choice comes from the fact that for many posts/tweets, we may not have images. So, for such cases, the system should be able to infer from just one modality. Since, it is a classification problem with imbalanced classes, we use F1 score to evaluate our model. The dataset used and results of the trained classifier is outlined in Table I and II respectively.

TABLE I
DATASET USED TO TRAIN FOOD/NON-FOOD CLASSIFIER

Dataset	# samples	# Food posts	# Non-Food posts
Train	44424	26912	17512
Val	5553	3364	2189
Test	5553	3364	2189

TABLE II
F1 SCORE OF FOOD/NON-FOOD CLASSIFICATION

Dataset	Image Only	Text Only	Multi-modal
Train	0.978	0.985	-
Test	0.971	0.988	0.995

$$FS = \gamma * image\_score + (1 - \gamma) * text\_score \quad (2)$$

3) Food category classification: After sifting quality FnB quotes, we tagged each post into categories. We considered 5 food categories: Bakery, Confectionery, Beverage, Dairy and Snacks. Such categorization is beneficial because it enables us to comprehend how trends evolve across different categories (beverage vs Snacks, for example) as well as how some categories influence the others (bakery to confectionery). As a result, it is crucial to understand the context with regard to category rather than viewing it as simply Food. Also, we frame this problem as a multi-label classification problem since a single post can belong to multiple categories. For example, a post stating 'Rainy Days call for chocolates And endless cups of Tea' should be categorized into Confectionery as well as Beverage.

One way to formalize the problem is to think about how we can represent posts in a manner that preserves good category properties, meaning that similar items should have similar representations. A natural representation in this case would be to use embeddings that preserve intuitive relationships

TABLE III
RESULTS OF FOOD CATEGORY CLASSIFICATION

Category	Dataset		Accuracy		F1 score	
	#samples in Trainset	#samples in Testset	Baseline SBert	Finetuned SBert	Baseline SBert	Finetuned SBert
Bakery	7303	3652	0.854	0.898	0.859	**0.909**
Snacks	1071	523	0.937	0.936	0.459	**0.496**
Dairy	2972	1408	0.867	0.891	0.637	**0.75**
Confectionery	1015	482	0.944	0.938	0.478	**0.494**
Beverage	2064	1046	0.931	0.931	0.753	**0.774**
Overall	14297	7043	0.893	0.919	0.70	**0.798**

between items. For example, we would expect that post that talks about 'Mango Cake' and 'Nutella Waffle' to be more similar to each other than to post that talks about 'Almond milkshake' because the former are both Bakery food items while Milkshake is a Beverage. It has become normal practice in NLP to optimize a large pre-trained model like BERT for a downstream task. The quality of the learnt metric, however, is highly influenced by the quality of the annotated training set that we do not have access to. For uncommon classes with little data samples, this issue is more severe. We get over this problem by automatically creating samples from unlabeled data and learning a representation for the label using a Siamese Neural Network architecture with Triplet loss [12].

We hand-curated small samples for each category and then generated triplets (anchor, positive and negative) from it. We used this dataset to fine tune Sentence Bert with model 'all-MiniLM-L6-v2' using triplet loss. The triplet loss seeks to push similar instances together and separate dissimilar examples in the latent space. As a result, distances between comparable samples are preserved when the encoders for the quotes are embedded into the same latent space. For the purpose of comparing the effectiveness of the triplet loss method, we experimented with 2 approaches: 1. Baseline using pre-trained Sentence Bert embedding 2. Triplet-loss based fine-tuned Sentence Bert embedding. We fine-tuned with batch size of 8 and Learning rate as 2e-5. We use the fine-tuned embeddings for multi-label text classification. Table III shows a significant improvement in multi-label classification by using triplet-loss fine-tuned SBert embedding. The category distribution samples are highly skewed with Bakery having the highest number of samples. However, using triplet loss based approach, we found that there has been noteworthy jump even for minority classes that helped spike the overall performance. This supports our approach of fine-tuning SBERT using Triplet loss by using a small hand-curated dataset.

C. Deduplication

While working with social data, we can find some posts that have exact/near duplicate content. To reduce the over-indexing on a few messages, we use a deduplication approach to apply a max cap on them. Most of the time such posts were from businesses trying to promote their product/event. Other times users just append emojis or make a small change to the text (Twitter Retweet, for example), so, doing an exact regex match doesn't solve the problem completely [14]. We extracted the embeddings of text_only part of each post using Sentence Transformer model 'nli-roberta-base-v2' [19] and then computed cosine similarities 4 between embeddings. Lastly, we consider the posts with $\geq$ 0.9 cosine similarities as duplicates.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

VI. INSPECTION AND EVALUATION

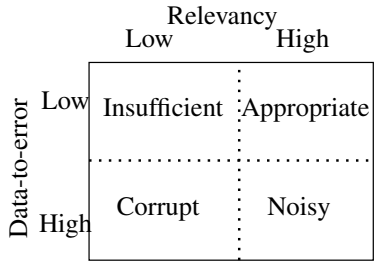

Fig. 3. Taxonomy for Social Data Quality
Taxonomy for Social Data Quality

The third phase of our framework is validation of sifted data. A formal investigation on the usefulness of this system on various downstream tasks like sentiment or trends analysis is the ideal way to evaluate the effectiveness of the framework. However, curating annotated data for such analysis is expensive and time consuming. Hence, through this work we introduce two metrics that we used to assess the data quality: data-to-error ratio and relevancy.

We got our expert Quality Check team to spot check the data points returned by our framework. This human-validated data is used to compute the data-to-error ratio as expressed in eq. 4.

$$Data - to - Error = \frac{\#samples\_marked\_as\_irrelevant}{Total\_samples\_validated} \quad (4)$$

The data to error ratio decreased drastically from 8.76% to 0.01% after the framework deployment. However, it is simple to determine if the social media postings that have been returned are inaccurate, but it is more challenging to determine how many posts that should have been returned but were

instead overlooked by the algorithm. So, we used relevancy as defined in eq. 5 as a metric check on how much data was accepted during the sifting process.

$$relevancy = \frac{\#samples\_marked\_relevant\_by\_framework}{Total\_samples\_collected} \tag{5}$$

We consider these two metrics together as an indicator of the quality of our collection and the sifting process as shown in Figure 3. The four Taxonomies that we used to describe data quality is defined as below:

- **Insufficient**: Most data points are valid but the number of data points returned are low. This can arise if framework has high False Negatives rate to have high precision i.e high Precision, low Recall
- **Appropriate**: Sufficient and valid data points that gives good estimates on various analytics, ie. high Precision and high recall
- **Corrupt**: When we have only a few data points and that is mostly unrelated to the study, this prohibits performing analytics or gives erratic results. This can arise if the framework returns only a few data points and those are mostly False Positives
- **Noisy**: This is the current state of the data, where the collected data has some invalid data points i.e False positives

So, the goal of our framework is to transform Noisy data into Appropriate Data. We could achieve a low data-to-error ratio (0.01%) with a relevancy score of 68.87% . Although it might seem a low relevancy score, but since we used many broad terms for data collection to get more data points, it is quite good. Thus, both metrics indicates strong evidence towards the appropriateness of data for analytics.

VII. Conclusion

Social media analytics have gained popularity in recent years, however many of the systems still can't deliver accurate data for practical applications. In this work, we've discussed the difficulties we faced when analyzing social data for FnB analytics. We've provided a framework: FOODIE for cleaning and preparing the data for analytics. Moreover, with easily annotatable data that is not highly particular to downstream tasks that call for task-specific domain specialists, two measures are introduced to estimate data quality. These generalized metrics can be very useful to assess data quality in many different scenarios where getting labeled data is either expensive or infeasible. We achieved more than 8% improvement in data quality by FOODIE framework and companies similar to us can achieve the same or more. This said, there is still scope in further improving the performance of ML based models like food category classification which will be the focus of our future work.

References

[1] Pyspellchecker, 2022. Accessed: 2022-09-15.

[2] Akankasha and Bhavna Arora. A review of sentimental analysis on social media application. In Vijay Singh Rathore, Marcel Worring, Durgesh Kumar Mishra, Amit Joshi, and Shikha Maheshwari, editors, *Emerging Trends in Expert Applications and Security*, pages 477–484, Singapore, 2019. Springer Singapore.

[3] Laura Barre, K Cronin, and A Thompson. What people post about food on social media. *Journal of Nutrition Education and Behavior*, 48(7):S52, 2016.

[4] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

[5] Rob Churchill and Lisa Singh. textprep: A text preprocessing toolkit for topic modeling on social media data. In *DATA*, pages 60–70, 2021.

[6] Nadine Côrte-Real, Pedro Ruivo, and Tiago Oliveira. Leveraging internet of things and big data analytics initiatives in european and american firms: Is data quality a way to extract business value? *Information & Management*, 57(1):103141, 2020.

[7] Jacob Danovitch. Linking social media posts to news with siamese transformers. *arXiv preprint arXiv:2001.03303*, 2020.

[8] Xuefan Dong and Ying Lian. A review of social media-based public opinion analyses: Challenges and recommendations. *Technology in Society*, 67:101724, 2021.

[9] Imane El Alaoui, Youssef Gahi, and Rochdi Messoussi. Big data quality metrics for sentiment analysis approaches. In *Proceedings of the 2019 International Conference on Big Data Engineering*, pages 36–43, 2019.

[10] Gal Ben David Grant Jenks, Javier Honduvilla Coto. Python word segmentation, 2018. Accessed: 2022-09-15.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pages 84–92. Springer, 2015.

[13] Heather Kennedy, Thilo Kunkel, and Daniel C Funk. Using predictive analytics to measure effectiveness of social media engagement: A digital measurement perspective. *Sport Marketing Quarterly*, 30(4), 2021.

[14] Aditya Kumar, Sneh Gupta, Ankit Sahu, and Mayank Kant. Deriving customer experience implicitly from social media. In *Companion Proceedings of the Web Conference 2022*, pages 131–135, 2022.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[16] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[17] Belal Abdullah Hezam Murshed, Jemal Abawajy, Suresha Mallappa, Mufeed Ahmed Naji Saif, Sumaia Mohammed Al-Ghuribi, and Fahd A Ghanem. Enhancing big social media data quality for use in short-text topic modeling. *IEEE Access*, 10:105328–105351, 2022.

[18] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[19] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[20] Camilla Salvatore, Silvia Biffignandi, and Annamaria Bianchi. Social media and twitter data quality for new social indicators. *Social Indicators Research*, 156(2):601–630, 2021.

[21] Ashutosh Singla, Lin Yuan, and Touradj Ebrahimi. Food/non-food image classification and food categorization using pre-trained googlenet model. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, pages 3–11, 2016.

[22] Chanin Songchon, Grant Wright, and Lindsay Beevers. Quality assessment of crowdsourced social media data for urban flood management. *Computers, Environment and Urban Systems*, 90:101690, 2021.

[23] Sonam Srivastava, Mahesh Kumar Singh, and Yogendra Narain Singh. Social media analytics: current trends and future prospects. In *Communication and Intelligent Systems*, pages 1005–1016. Springer, 2021.

[24] Dandan Tao, Pengkun Yang, and Hao Feng. Utilization of text mining as a big data analysis tool for food science and nutrition. *Comprehensive reviews in food science and food safety*, 19(2):875–894, 2020.