Iterative Approach for Novel Entity Recognition of Foods in Social Media Messages

Brandon Chenze
Department of Computer Science
California State University, Fullerton
Fullerton, California 92831, USA
bchenze@csu.fullerton.edu

Eugene Lee
Department of Computer Science
California State University, Fullerton
Fullerton, California 92831, USA
eugene097@csu.fullerton.edu

Anand Panangadan
Department of Computer Science
California State University, Fullerton
Fullerton, California 92831, USA
apanangadan@fullerton.edu

***Abstract*—Entity recognition is the computational task of identifying words or phrases in natural language text that correspond to real-world objects of specific predefined types and has several text processing applications. However, current entity recognition methods are trained to recognize only a relatively small set of entity types. Extending an entity recognition method to a novel entity type requires a large labeled dataset of known mentions of the new entity type. As labeling natural language datasets is a time-consuming process, identifying novel entity types remains a challenging problem. This work extends the Snowball approach to enable recognition of novel entity types from unstructured text that is typical in social media. The approach uses a set of keywords known to be associated with a new entity type and a large unlabeled corpus of text that could contain mentions of the entities. The iterative approach starts with dataset messages that are most likely to contain the entities. Likelihood is based on the number of keywords that appear in a message. This approach is then applied to the problem of identifying food entities in messages on the Twitter network. The initial set of keywords is obtained from the FoodKeeper dataset, a dataset provided by the U.S. Food Safety and Inspection Service, and which contains information on a variety of foods. The motivation for this application is to build a system that can automatically respond to messages about food with relevant information about food safety and preparation in an effort to reduce food waste. We evaluated the precision and recall of the entity recognition method on a hand-labeled dataset of tweets. The system achieved a precision of 0.80 and a recall of 0.80 (f-score of 0.80) on this dataset.**

***Index Terms*—machine learning, named entity recognition, food waste, sustainability, FoodKeeper**

I. Introduction

Entity recognition is the computational task of identifying words or phrases in natural language text that correspond to real-world objects of specific predefined types. Entity recognition is a frequently used component in many text processing applications, including question answering, text summarization, and machine translation [1]. Entity recognition capability is available in popular natural language processing software libraries such as NLTK [2]. However, a particular implementation is capable of recognizing only a relatively small set of entity types such as *person*, *location*, and *organization* [3]. This limitation arises because entity recognition is based on machine learning and therefore requires a large labeled dataset to identify the patterns corresponding to a particular entity type. Entity recognition libraries such as SpaCy give the end-user the ability to train the system to recognize new entity types provided sufficient labeled data is provided [4]. However, labeling natural language datasets is a time-consuming process and hence this ability is of limited use.

In our prior work [5], we described how food-related entities can be recognized using the SpaCy system with only simple keyword matching using a known set of food-related words. This baseline approach had high precision (0.96) but low recall (0.52). In this work, we describe how a "Snowball" approach can be used to improve the recall of learning of new entity types from unlabeled social media datasets, starting from a set of known keywords associated with that entity type. This method is an iterative approach, where in every iteration the system learns entity occurrence patterns that match a set of representative entity mentions and then extracts new mentions using the previously learnt patterns. The approach can make use of any entity recognition learning algorithm - the iterative portion determines the training data provided to the recognition algorithm. This method was first described for extracting relations from webpages in the DIPRE [6] and Snowball [7] systems. (The name "snowball" refers to the increasing number of the entities recognized in every iteration, analogous to how a snowball increases in size as it rolls down.) The basic Snowball approach is most accurate when the relevant entity mentions appear within consistent contexts as patterns are matched to sentences using exact or approximate string matching (DIPRE and Snowball, respectively). Thus, directly applying the basic Snowball method to social media messages will not succeed since these messages are typically short and use language in idiosyncratic ways. For instance, capitalization is an important feature for named entity extraction, but this feature is used inconsistently in Twitter where words are often capitalized for emphasis, and named entities can often appear in lowercase [8].

In this work, we extend the Snowball approach to enable recognition of novel entity types from unstructured text that is typical in social media. Specifically, we are given a set of keywords known to be associated with that entity type. Note that mere appearance of a keyword is not necessarily a valid entity since in natural language a word can have different meanings. We start the iterations with dataset messages where

we have the highest confidence that the known keywords represent true entities. Confidence is based on the number of keywords that appear in a message. Intuitively, we expect that a message that has multiple keywords is more likely to be using these keywords in the same context. An overview of the different steps in the approach is shown in Fig. 1.

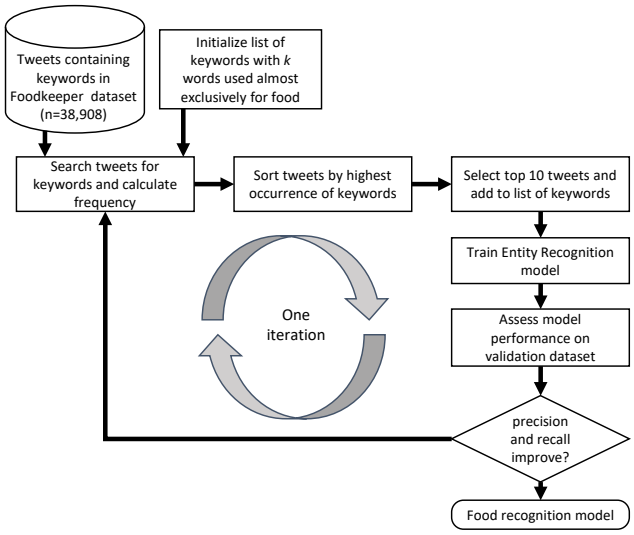

Fig. 1. Steps in the iterative entity recognition pipeline.

We applied this approach to the problem of identifying food entities in messages on the Twitter network. The set of keywords is available from the FoodKeeper dataset, provided by the U.S. Food Safety and Inspection Service [9]. This dataset contains information on a variety of foods and for each type of food, how long it remains safe for consumption, typical methods of preparation of that food item, and how to ensure food safety. The motivation for this application is to build a system that can automatically respond to messages about food with relevant information about food storage in an effort to reduce food waste.

We evaluated the precision and recall of the Snowball method on a hand-labeled dataset of tweets. The system achieved a precision of 0.80 and a recall of 0.80 (f-score of 0.80). For comparison, our previous non-iterative approach resulted in a model that achieved a precision of 0.96, a recall of 0.52, and f1-score of 0.68 [5]. Thus, the inclusion of the Snowball technique improves the recall by a large amount but at the cost of lower precision.

The main contributions of this work are (1) an extension of the Snowball approach to recognize mentions of new entity types in short social media messages from unlabeled data, and (2) application of the proposed method for detecting food entities in the Twitter network.

The rest of the paper is organized as follows. Section II lists related work in entity recognition, particularly on learning to recognize new entity types. Section III gives detailed information on the datasets used in this work. Section IV describes the steps of the approach. We present our results in Section V and give our conclusions in Section VI.

II. Related Work

Entity recognition methods have been extensively studied for natural language text datasets, but mainly for identifying all instances of a predefined set of entities [1]. These methods make use of natural language parsing and entity-specific features learned from a labeled dataset.

Brin [6] introduced the DIPRE method to extract instances of a specific type of relation (such as author-book) from a collection of webpages using only a small representative set of instances of that relation. Agichtein and Gravano extended this method by using approximate matching instead of exact pattern matching [7]. These early methods used little or no syntactic information. More recent approaches add syntactic information based on parsing the input sentences. Yates et al. [10] presented TEXTRUNNER, an open domain relation extraction system for the Web. TEXTRUNNER uses a natural language parser during the training phase. These approaches generally work well when the input data is either well-formed natural language sentences or when the context surrounding the relation instances is consistent (as in the case of webpages generated from a database).

We are instead interested in social media networks where natural language is used in idiosyncratic ways and with high variability. Doan et al. [11] summarize the issues in information extraction from unstructured data including that of representing the extracted information and managing evolving content. Sakaki et al. [12] describe a method to detect earthquake-related tweets. The method uses features specific to this application. Benson et al. [13] used machine learning to identify artists and venues from tweets. It develops a graphical model by learning records and records-message alignment. Ritter et al. [8] use latent variable modeling to extract event types described in tweets. Features such as tweet popularity and the times of events referred to in the tweets are also used. Zhao et al. [14] describe a method to extract only the most "topical" keywords from tweets. Our approach differs from these methods in that the novel entity type is given only as a set of terms describing representative entities. Thus, arbitrary types of entities can be discovered without utilizing type-specific features.

The following works are designed for messages on social media networks. Lu et al. [15] uses text snippets from Reddit to identify linguistic features that predict the likelihood of transitioning from casual drug discussion forums to drug recovery forums. Ritter et al. [16] present a processing pipeline that includes part-of-speech tagging, chunking, and named entity recognition to improve performance on short text (tweets) compared to traditional named entity recognition systems such as Stanford NER. Their work uses LabeledLDA with Freebase dictionaries in contrast to the SpaCy approach that uses the Transformer architecture. Zhao et al. [17] uses Twitter-LDA, an unsupervised topic modeling approach, to discover topics from Twitter. Their work is on topic

discovery, which is not necessary for identifying a novel entity, since we already know the "topics" in this application. Ling and Weld [3] represent entities as a set of tags; this enables a much larger number of entities to be discovered but is still restrictive if the entity of interest does not match the predefined set of tags. Lin and Pantel [18] also describe an unsupervised algorithm to generate a general set of inference rules from text documents.

III. Datasets

A. Twitter Dataset

We used a large (1.6 million) publicly-available collection of tweets from 2009 for the study [19]. Each record in the dataset has the following attributes: the tweet identification number, tweet creation timestamp, whether the data was queried, the Twitter user that posted the tweet, and the text content of the tweet. In this study, we used only the text content of the tweets (i.e., not any of the meta-data). The dataset has a large variety of tweets. Specifically, in the context of this paper, the dataset includes both food-related and tweets unrelated to food. The large number of tweets and wide variety of tweets in the dataset enable us to use different subset for both training and validation sets for entity recognition.

B. FoodKeeper Dataset

To initialize the iterative approach to identify food-related words in a tweet, we make use of the FoodKeeper dataset provided by the U.S. Food Safety and Inspection Service [9]. This dataset contains information about 509 different foods and their associated categories as well as tips on the sustainable use of each product. The data is categorized into four sections: the categories of food items, relevant storage methods for the different foods, tips for cooking specific products, and the various methods which are commonly used to cook the food products.

The product section contains a variety of sustainable use information on different food products. Along with the product name, the dataset includes the commonly associated keywords for that product. In this study, we use only the product information contained in the dataset. The information in the other sections could be used to develop social media-based applications to discourage food waste using the food entity recognition approach described in this paper.

IV. Approach

A. Training Set Creation

Named Entity Recognition typically uses supervised machine learning and therefore requires a training dataset which informs the model of what is correct versus incorrect. A dataset which correctly labels tweets as food-related or not was not unavailable and therefore we created such a dataset. The methodology for creating this was to first generate a list of the unique foods contained in the FoodKeeper dataset. Once the unique food keywords were gathered, they were used as a search query on Twitter's API to gather tweets that we considered to be related to food. A training set of over 40,000 tweets were gathered for the 509 unique foods. This method was expected to yield a high recall but low precision model due to the informal nature of text on social media; however in practice our findings proved otherwise. Furthermore, our iterative method described below further improved on our initial results.

B. Pre-processing

Multiple pre-processing steps were performed on the tweets prior to model creation. Some steps, along with our justification for them, are as follows:

1) **Convert to lowercase:** Entities should be recognized regardless of their capitalization. Although removing case removes some information that could be useful for other language tasks such as sentiment analysis, for our purposes, this was not necessary.
2) **URL replacement:** Twitter usernames and URLs are replaced by the tags <USERNAME>and <URL>, respectively. The purpose of removing URLs and usernames is to prevent the model from mistaking URL patterns as food entities due to the high number of occurrences throughout the training data.
3) **Remove punctuation:** Sometimes food words occur at the end of a sentence followed with punctuation without a space. To prevent unintended exclusion of food entities, we removed punctuation.

C. Natural Language Processing

A series of natural language processing steps were performed to transform the instances in the training data into feature vectors for the named entity recognition task. These steps were organized into an NLP pipeline with use of the open-source NLP library called spaCy[1]. Specifically, we make use of spaCy v3 Named Entity Recognition model which is based on the transformer network architecture. This architecture is composed of 2 parts: Tok2Vec and the transformer itself.The Tok2Vec applies a "token-to-vector" model. The use of word embeddings helps capture the semantic meanings of words by creating a feature vector representation of a word. This can then be shared between the DependencyParser, Tagger, and EntityRecognizer components. Utilizing the training set described in the previous section, we trained this entity recognizer to work as a binary classifier to differentiate between food entities like "cheese" or "chicken" and non-food entities like "paint" or "C++". We represented the training data in the widely used Inside-Outside-Beginning text tagging format. The format contains the original tweet, the food entity contained in the tweet, and the start and stop positions of the entity words.

D. Snowball method

The Snowball method is a process used during the training of a machine learning model where multiple models are

[1]https://spacy.io/

trained over multiple iterations. With this approach, the first iteration uses a small but precise pool of training data to create a model which can then be used to identify *new* data to retrain the model in the following iteration. The initial small data pool will ensure the model is accurate but not overly broad that it identifies irrelevant entities. Due to the iterative nature of this process, starting with inaccurate or overly general keywords greatly reduces the precision of models created in successive iterations. Therefore, starting with a small but accurate set of keywords helps maintain the model accuracy while also allowing the model to identify broader categories of entities with more certainty from the context gained in each iteration. This method is used in the process of training our Named Entity Recognition model to achieve a model that can identify, with high degree of confidence, the different food entities within a tweet. The steps in the iterative approach are illustrated in Fig. 1. The approach begins with a small pool of keywords that are highly likely to be related to the topic of food. To determine such keywords, each tweet was analyzed to count the occurrences of each word that exists within the Name column on the Product page within the FoodKeeper dataset. The top words are then used as the initial keywords to begin the Snowball process. The exact count of initial keywords is a hyperparameter and leads to different results in the model. Once the initial keywords are found, each tweet is then analyzed and ranked to determine if it will be used for training in the next iteration. Initially, the ranking of a tweet is found by counting the number of keywords that matched within the tweet; after the first iteration the previous iteration's model is used to find the food entities based on a ranking mechanism. These tweets are then used as training data to create an initial model. Once the initial model is created, the training tweets are then analyzed again to find new keywords by cross-referencing the FoodKeeper database. This process repeats until no new keywords are found. At this point, the loop lowers the ranking threshold to allow for more training data at the cost of less assurance of proper context. Once no more new keywords are found and the ranking requirement is less than 1, then the training loop ends, and the final model is available to use.

V. Results and Discussion

Prior to the iterative Snowball method introduced in this paper, our baseline model achieved 73% test-set accuracy with a precision of 0.96, a recall of 0.52 and f1-score of 0.68 on a hand-labelled test set of randomly collected tweets [5]. This is similar to the accuracy of the system described by Lu et al. [15] for identifying drug-related terms on the Reddit social media platform. Their system achieved an accuracy of 69.3% on a test dataset and 82.3% accuracy with tuning.

The Snowball approach was tested with different values of its hyperparameters. The hyperparameters that we tuned were the *rank* required for a tweet to be accepted to be used to train the model, and the number of initial keywords to start the iterative process. The approach is sensitive to the specific values of these two hyperparameters. If the rank for tweets is too high, the training loop will find very few tweets. Conversely, if the rank is too low, then the model may learn from tweets that are not related to the topic resulting in a model that predicts a high number of false positives. Similarly, if the initial keyword count is too small, the process is unable to find a sufficient number of tweets to train the model. On the other hand, if too many keywords were used, the Snowball method would be completed in a few iterations, insufficient to fully gather context.

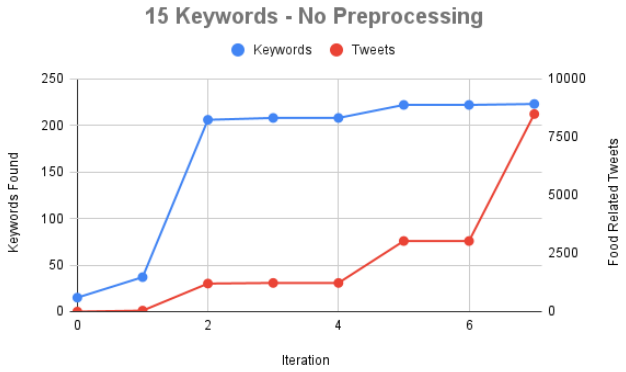

Fig. 2. Number of keywords and food-related tweets found per iteration with no pre-processing of training data.

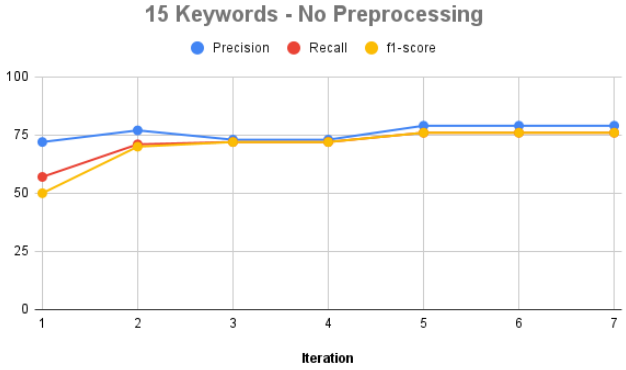

Fig. 3. Precision, recall, and f1-score at every iteration with no pre-processing of training data.

We then evaluated the performance of the system with different levels of pre-processing applied to the training data.

A. Evaluation with no pre-processing

For the first set of evaluations, no pre-processing was performed on the tweets and they were left in their original state. The number of initial keywords was set to 15 and the tweet rank for acceptance was set to 3. Fig. 2 shows the number of keywords and the number of food-related tweets identified in every iteration. The training loop executed a total of 9 iterations finding a total of 223 keywords and 8,488 tweets. The change in precision, recall, and f1-score in every iteration is shown in Fig. 3. The final model had a precision of 0.79, a recall of 0.76, and an f1-score of 0.76.

B. Evaluation with Twitter-specific pre-processing

Next, the tweets were pre-processed by replacing Twitter usernames and URLs with <USERNAME>and <URL>tags and converting the text into lowercase. However, punctuation was left unaltered. The initial keyword count was 15 and the rank for acceptance was 3.

Fig. 4 shows the number of keywords and the number of food-related tweets identified in every iteration. The method found the same 223 keywords among 8,488 tweets but in 8 iterations. The change in precision, recall, and f1-score in every iteration is shown in Fig. 5. The final model had a precision of 0.79, a recall of 0.76, and an f1-score of 0.76.

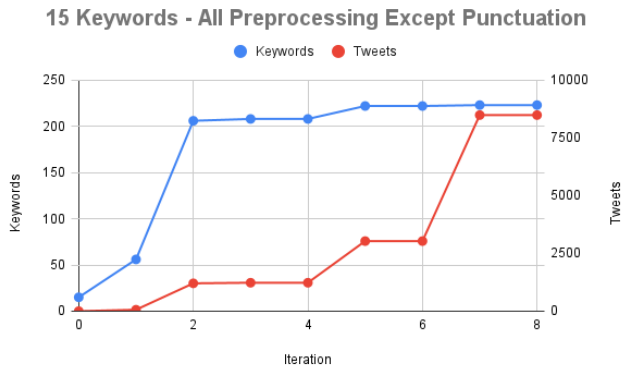

Fig. 4. Number of keywords and food-related tweets found per iteration with Twitter-specific pre-processing applied to the training data.

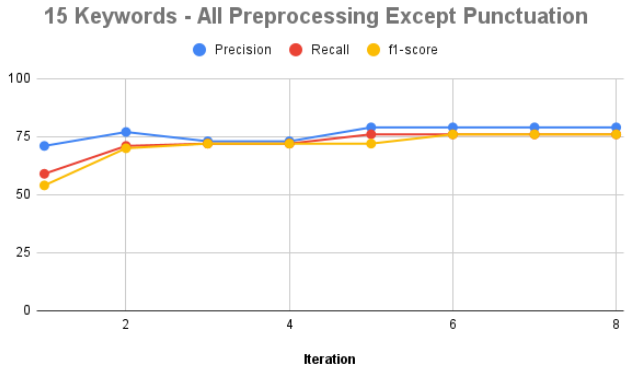

Fig. 5. Precision, recall, and f1-score at every iteration with Twitter-specific pre-processing applied to the training data.

C. Evaluation with extensive text pre-processing

For the next set of experiments, the tweets were pre-processed by removing all punctuation (periods, hyphens, and commas), replacing twitter usernames and URLs with <USERNAME>and <URL>tags, respectively, and converting the text into lowercase. The initial keyword count was set to 15 and the rank for acceptance was set to 3.

Fig. 6 shows the number of keywords and the number of food-related tweets identified in every iteration. The training loop executed a total of 8 iterations finding 216 keywords and 4,108 tweets. The corresponding change in precision, recall, and f1-score at every iteration is shown in Fig. 7. The model has a precision of 0.8, recall of 0.8, and f1-score of 0.8.

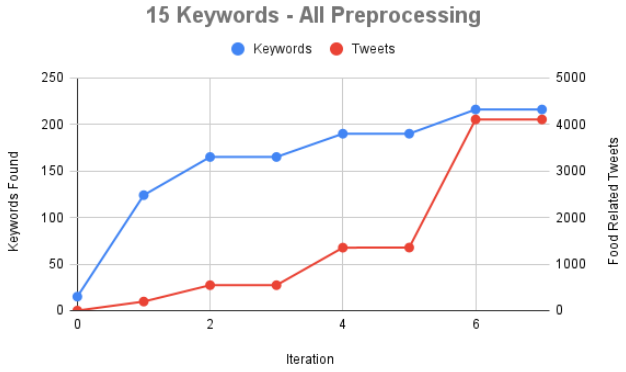

Fig. 6. Number of keywords and food-related tweets found per iteration with all pre-processing steps applied to the training data.

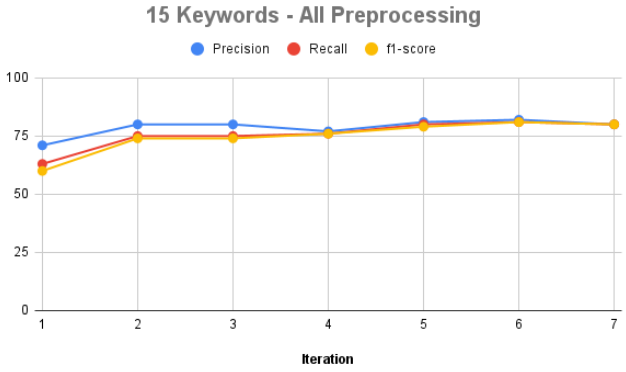

Fig. 7. Precision, recall, and f1-score at every iteration with all pre-processing steps applied to the training data.

Additional tests were also performed to gauge how well food-related entities would be identified in real-world tweets by the different models. A sample of tweets was randomly selected from the Twitter dataset for evaluation and the entities recognized by the different models were compared. Fig. 8 shows example food entity identification results using the model created from tweets with extensive pre-processing.

VI. Conclusions and Future Work

We described the application of the Snowball approach to novel entity recognition that can be used to identify food entities from messages in the Twitter network. The results from evaluation indicate that the inclusion of the Snowball technique during the training of a model improves the recall and f1-score at the cost of a lower precision. Our previous non-iterative approach resulted in a model that achieved a precision of 0.96, a recall of 0.52 and f1-score of 0.68 [5]. In comparison, the best test when using the snowball method

@ellelovexx haaaaa i want mac & cheese toooooo!!! hahahaha hey..i still got the one u left here...i guess im making that today Oo lol

@Sabbyaz aiyooooo maybe chocolate will help? chocolate food helps in most situations

Not wanting to get rid of her rabbits. This is going to be a great day..

@Jonasbrothers thats so exciting! u are coming to south america in a few hours! but not to Colombia hope u have fun here! we love u!!

@rawrmtoxic i have too much veggies and rice. I WANT PIZZA BUT NO. RECESSION.

@cuteredshoes GREAT! now i want chocolate .

Thought Adventure Land was good., not as good as Superbad

Zzzzz....packing and cleaning all day. Pizza and glasses of red wine....now...paint the 4th room....ggrrrrr, don't want to, but I must.

@_Jaska Some things... they just never get old. http://tinyurl.com/holdisgiantcherry I miss Maya.

Fig. 8. Example food entities (shaded in gray) identified using the model created from tweets with extensive pre-processing.

achieved a precision of 0.80, a recall of 0.80, and f1-score of 0.80. Thus, using the Snowball approach resulted in a significant increase of 0.28 in recall performance. Thus, through the inclusion of the Snowball method, the likelihood that a food entity is identified from the tweet is high while still limiting amounts of false positives.

As a future application, this entity recognition model can be used to determine whether tweets are food-related. In the case that they are food-related, an insightful response can automatically be generated using information from authoritative sources (such as USDA's FoodKeeper dataset) and posted as a reply, informing the poster of different ways of maximizing the longevity of their specific food.

ACKNOWLEDGMENT

This work is supported by Hispanic Serving Institutions Education Grants Program grant no. 2019-38422-30211 from the USDA National Institute of Food and Agriculture.

REFERENCES

[1] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[2] E. Loper and S. Bird, "Nltk: The natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[3] X. Ling and D. S. Weld, "Fine-grained entity recognition," in *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[4] X. Schmitt, S. Kubler, J. Robert, M. Papadakis, and Y. LeTraon, "A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 2019, pp. 338–343.

[5] E. Lee, B. Chenze, and A. Panangadan, "Encouraging sustainability practices through entity recognition of food items on social media," in *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2021, pp. 263–266.

[6] S. Brin, "Extracting patterns and relations from the world wide web," in *International workshop on the world wide web and databases*. Springer, 1998, pp. 172–183.

[7] E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plain-text collections," in *Proceedings of the fifth ACM conference on Digital libraries*, 2000, pp. 85–94.

[8] A. Ritter, O. Etzioni, and S. Clark, "Open domain event extraction from twitter," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 1104–1112.

[9] Z. Qamar, "Foodkeeper," *Journal of Nutrition Education and Behavior*, vol. 50, no. 1, p. 101, 2018.

[10] A. Yates, M. Banko, M. Broadhead, M. J. Cafarella, O. Etzioni, and S. Soderland, "Textrunner: open information extraction on the web," in *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2007, pp. 25–26.

[11] A. Doan, J. F. Naughton, R. Ramakrishnan, A. Baid, X. Chai, F. Chen, T. Chen, E. Chu, P. DeRose, B. Gao *et al.*, "Information extraction challenges in managing unstructured data," *ACM SIGMOD Record*, vol. 37, no. 4, pp. 14–20, 2009.

[12] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 851–860.

[13] E. Benson, A. Haghighi, and R. Barzilay, "Event discovery in social media feeds," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2011.

[14] W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achanauparp, E.-P. Lim, and X. Li, "Topical keyphrase extraction from Twitter," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 379–388.

[15] J. Lu, S. Sridhar, R. Pandey, M. A. Hasan, and G. Mohler, "Investigate transitions into drug addiction through text mining of Reddit data," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2367–2375.

[16] A. Ritter, S. Clark, O. Etzioni *et al.*, "Named entity recognition in tweets: an experimental study," in *Proceedings of the 2011 conference on empirical methods in natural language processing*, 2011, pp. 1524–1534.

[17] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing Twitter and traditional media using topic models," in *European conference on information retrieval*. Springer, 2011, pp. 338–349.

[18] D. Lin and P. Pantel, "Discovery of inference rules for question-answering," *Natural Language Engineering*, vol. 7, no. 4, pp. 343–360, 2001.

[19] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N project report, Stanford*, vol. 1, no. 12, p. 2009, 2009.