

**NATURAL LANGUAGE PROCESSING**

**SWE1017**

**RESTAURANT REVIEW ANALYSIS**

**J COMPONENT REPORT**

**DONE BY**

# HARINI GOKULRAM NAIDU - 19MIA1004 P SUBHASHRI - 19MIA1008

# Submitted to:

# Prof. Ranganathan Sridhar

**ABSTRACT:**

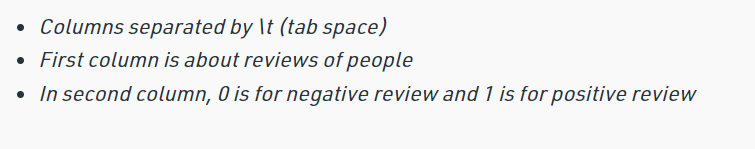
Restaurants nowadays prefer taking online orders. It not only helps in getting effective customer feedback but also useful for managing orders easily. We are moving towards an automated and digital world. Having a significant online presence is necessary for any restaurant to be successful and prosperous. Getting customer feedback and analyzing them in an effective manner makes the difference. This study analyses the restaurant reviews and presents useful information that the ratings do not consider or overlook. First we perform tokenization, stemming etc on the reviews and then summarization is done using gensim and results are displayed using effective visualization techniques. Machine learning algorithms like Naïve Bayes, Logistic regression, MLP, SVM, Decision tree and Random classifier is used for classifying the reviews in proper aspects. Future work is also discussed so that an efficient analysis system can be developed utilizing the potential of reviews.

**INTODUCTION:**

For years food and hospitality businesses are running on the assumption that good food and service is the way to attract more customers. But the advent of science and technology, more importantly, the data created by the use of online platforms has pointed towards new findings and opened new doors: Most consumers nowadays rate a product online, over 1/3rd of them write reviews and nearly 88% of the people trust online reviews. Review Services like Yelp, Google Reviews, etc. provide customers and businesses a way to interact with one another. Reviews and Ratings are useful sources of information but significant problems exist in extracting relevant information and predicting the future through analysis and correlation of the existing data. Each day thousands of restaurants and businesses are reviewed by the customers.

The main objective of the work proposed in this paper is to enhance the user experience by analyzing the reviews of restaurants and categorize them in some aspects so that a user can easily know about the restaurant. Restaurants are not able to utilize reviews for their businesses. We want to use the aspects that are important in the food and service industry so that we can analyze the sentiment of text reviews and help them to improve their businesses. This research paper work can be applied to many other industries related to food and hospitality.

**DATASET DESCRIPTION:**



**LITERATURE SURVEY**

2.1. A survey of Sentiment Analysis Challenges. The consequences of challenges in the area of sentiment analysis has been discussed. Sentiment review structure is compared with sentiment analysis challenges in the first distinction. The effect of this distinction shows that domain-dependence is an important part of sentiment challenges. The second comparison deals with the accuracy of sentiment analysis models based on the challenges. Structured , Semi-structured , and Unstructured are three types of review structures that were used for the first comparison. Theoretical and technological are the two types of sentiment analysis challenges. The challenges include Domain dependence, negation , bipolar words ,entity feature/ keyword extraction ,spam, or fake review, NLP overheads like (short abbreviations, ambiguity, emotions, sarcasm). Parts-of-speech (POS) tagging gives highly accurate results for the theoretical types of challenges. The phrases and expressions of n-gram give it an edge over all other techniques used for a technical set of challenges. The results explained the effectiveness of sentiment analysis challenges for improving the accuracy of the model .

2.2. Aspect based Sentiment Oriented Summarization of Hotel Reviews Due to the unstable size of review dimensions and customer produced content, different text analytic approaches like opinion mining , sentiment analysis, topic modeling , aspect classification, play a significant role in analyzing the content. Topic Modelling can find diverse topics in a corpus of text because of its statistical nature. For every aspect type, there is a certain opinion linked to it and the Sentiment analysis method can effectively bring out these emotions. Whether it is a business intelligence problem or a case of unstructured document categorization sentiment analysis is useful for most of the cases. It has emerged as the most important aspect of the Information Retrieval process. The strategies regarding text summarization can boost sentiment analysis research. The opinion mining of the hotel reviews is done using SentiWord library. The reviews were summarized on different aspects and sentiment analysis was performed.

2.3. Assessing the Helpfulness of Online Hotel Reviews.A review can be considered useful for decision making purposes only when it is thoughtful and insightful. The indicators representing the importance of reviews are different for diverse research areas due to ease of access. In the case of travel and hospitality websites, the reviews containing maximum votes are considered to be more informative and useful for consumers. It can be helpful in optimizing the cost of the search for most of the consumers by using feature engineering.

2.4. A framework for Fake Review Detection in Online. This paper paves an idea of the challenges that we could face in our research. Specifying the significance of online feedback for different types of industries and the amount of difficulty attached in procuring and maintaining a favorable honor on the Internet, diverse methods have been used to enhance digital existence, including unethical practices. Fake reviews are one of the most preferred unethical methods which exist on sites such as Yelp or TripAdvisor. In response to that Fake Feature Framework (F3), helps to assemble and constitute features for fake reviews choice. F3 estimates knowledge obtained from both the user (personal profile, reviewing activity, trusting information.

**METHODOLGY:**

**IMPORTING DATASET**

Import dataset with setting delimiter as ‘\t’ as columns are separated as tab space. Reviews and their category(0 or 1) are not separated by any other symbol but with tab space as most of the other symbols are is the review (like $ for the price, ….!, etc) and the algorithm might use them as a delimiter, which will lead to strange behavior (like errors, weird output) in output.

**TEXT CLEANING OR PREPROCESSING**

* **Remove Punctuations, Numbers**: Punctuations, Numbers don’t help much in processing the given text, if included, they will just increase the size of a bag of words that we will create as the last step and decrease the efficiency of an algorithm.
* **Stemming**: Take roots of the word
* **Convert each word into its lower case**: For example, it is useless to have some words in different cases (eg ‘good’ and ‘GOOD’).

**TOKENIZATION**

Tokenization involves splitting sentences and words from the body of the text.

 Making the bag of words via sparse matrix

* Take all the different words of reviews in the dataset without repeating of words.
* One column for each word, therefore there is going to be many columns.
* Rows are reviews
* If a word is there in the row of a dataset of reviews, then the count of the word will be there in the row of a bag of words under the column of the word.

We need CountVectorizer class from sklearn.feature\_extraction.text.   
We can also set a max number of features. Do the training on the corpus and then apply the same transformation to the corpus “.fit\_transform(corpus)” and then convert it into an array. If the review is positive or negative that answer is in the second column of the dataset.

**ML MODEL**

* Splitting Corpus into Training and Test set. For this, we need class train\_test\_split from sklearn.cross\_validation. Split is made as 80/20.  
  X is the bag of words, y is 0 or 1 (positive or negative).
* Machine learning models:

1. NAÏVE BAYES:

Naive Bayes is a machine learning model that is used for large volumes of data, even if you are working with data that has millions of data records the recommended approach is Naive Bayes. It gives very good results when it comes to NLP tasks such as sentimental analysis.

1. RANDOM FOREST

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm.

1. DECISION TREE

The decision tree classifier (Pang-Ning et al., 2006) creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute.

1. LOGISTIC REGRESSION

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

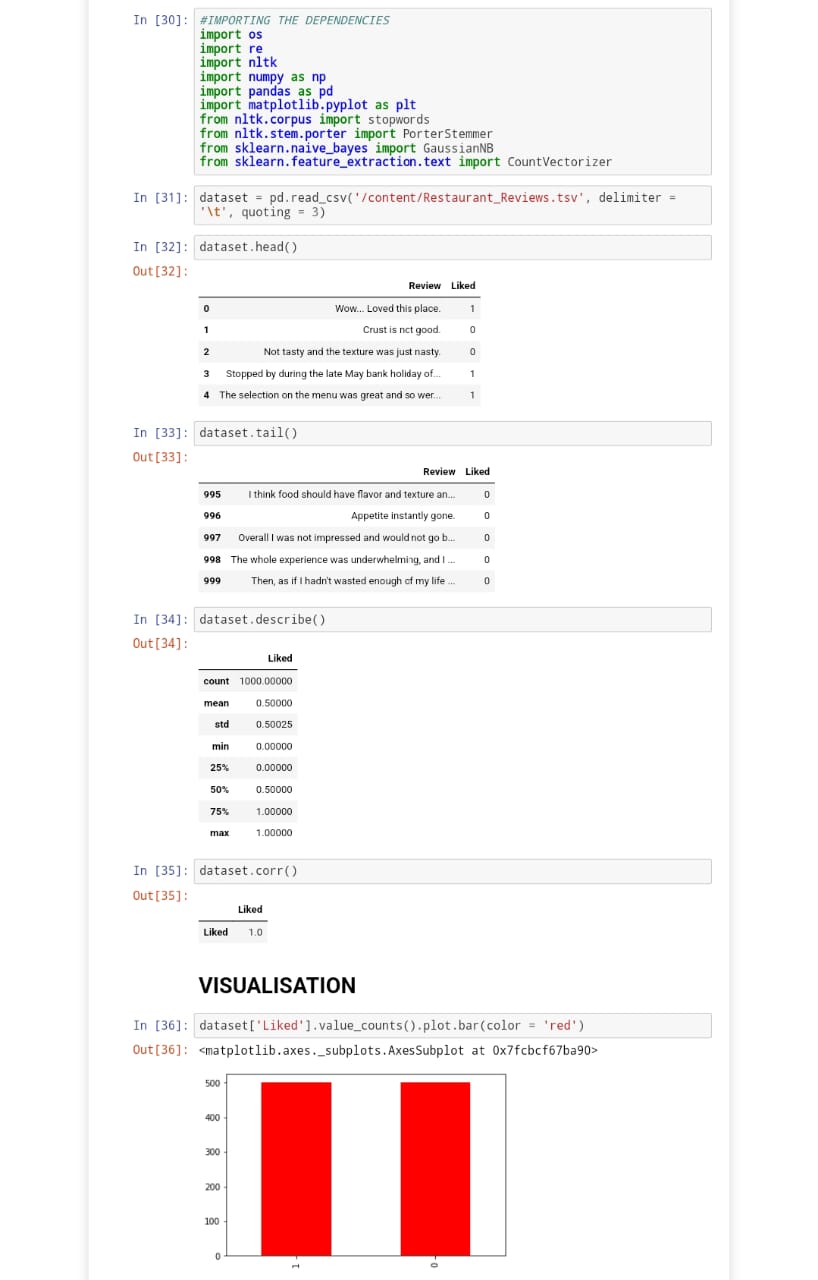
1. SVM

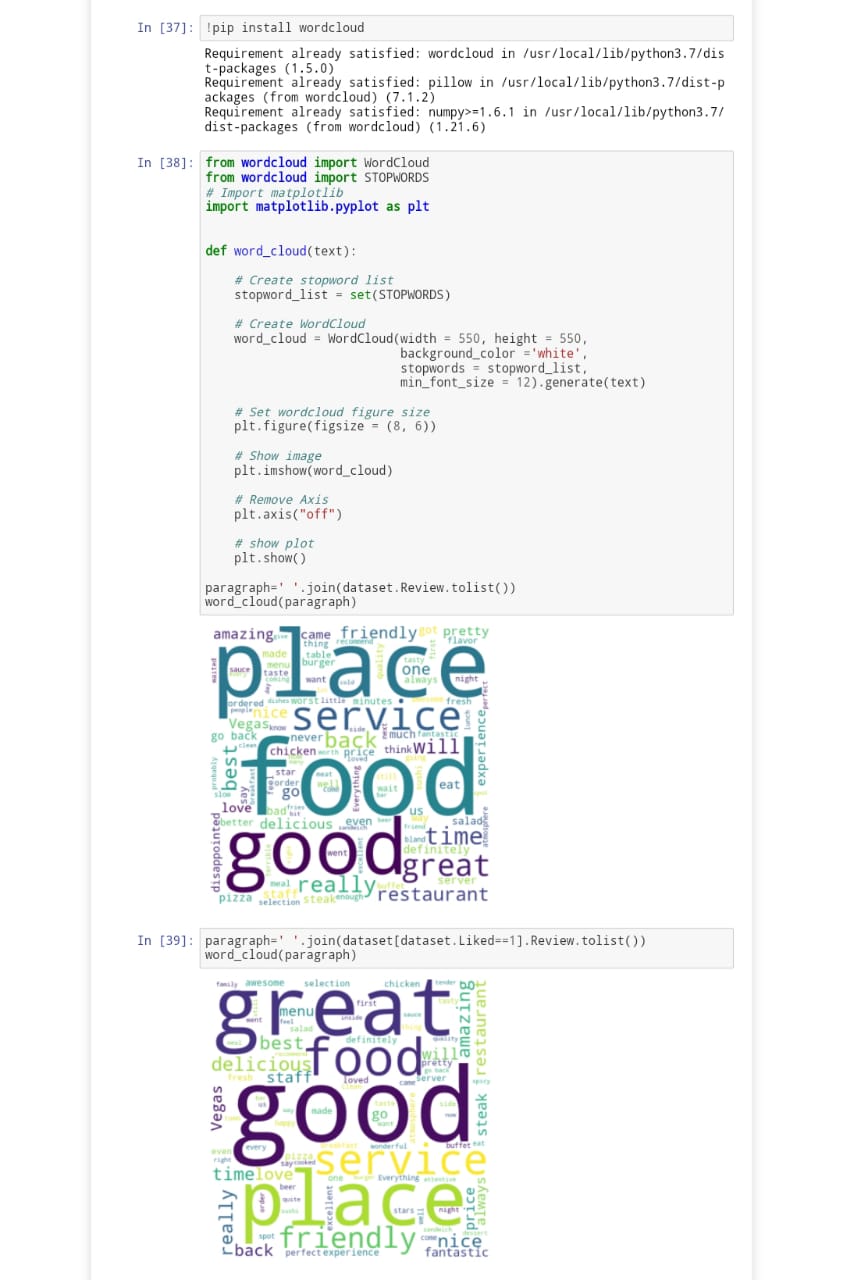
SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs.

1. MLP

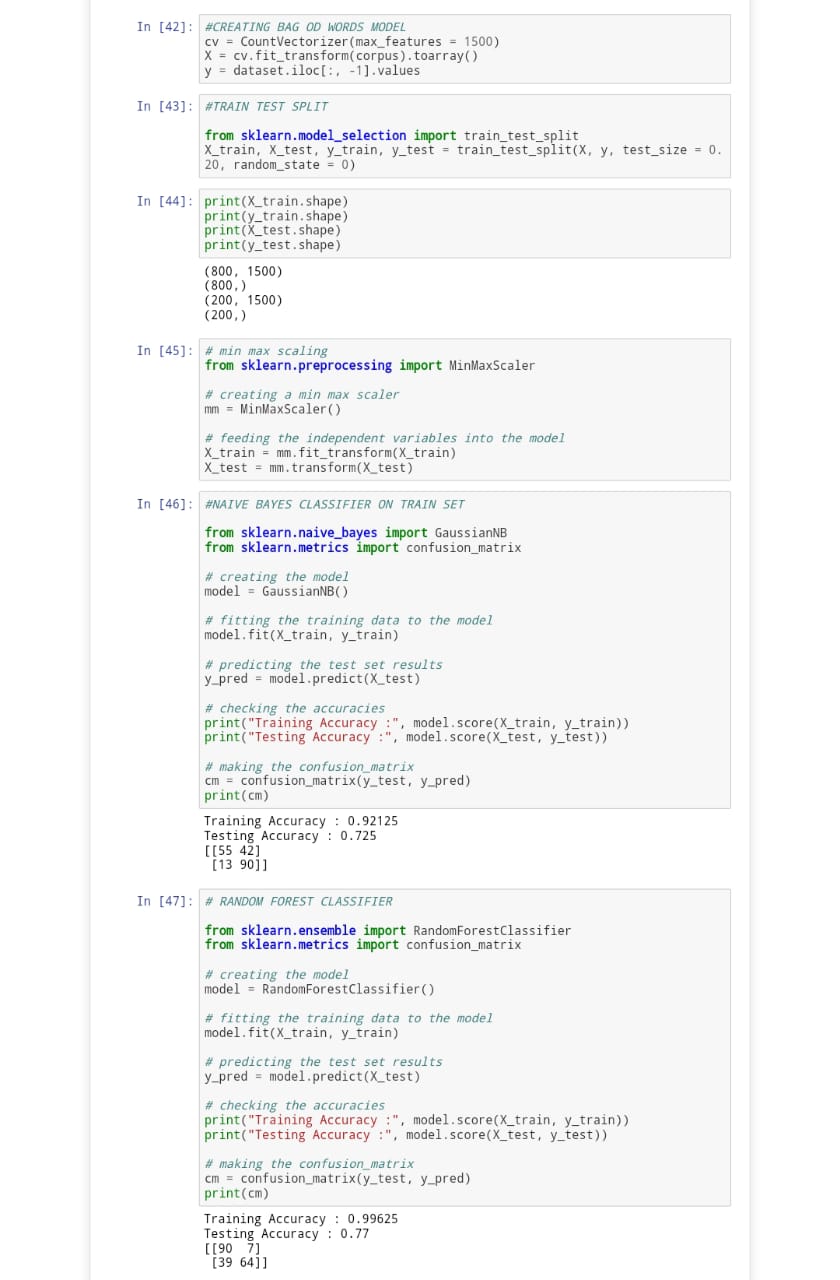
Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function f ( ⋅ ) : R m → R o by training on a dataset, where is the number of dimensions for input and is the number of dimensions for output.

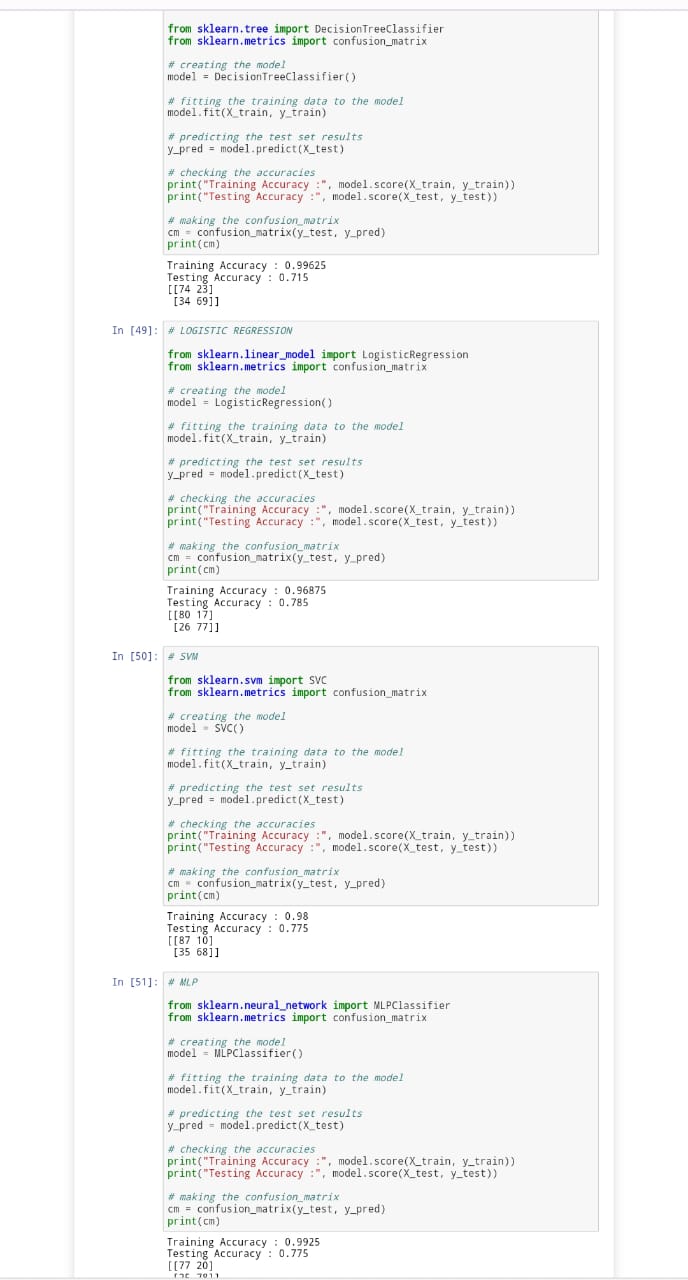
**IMPLEMENTATION**





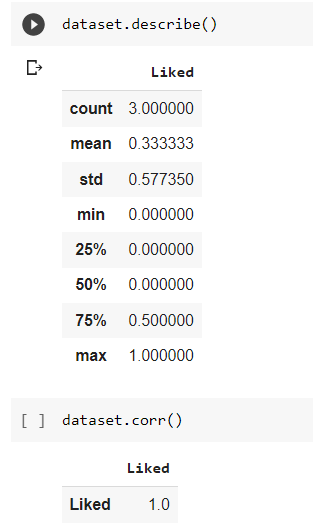


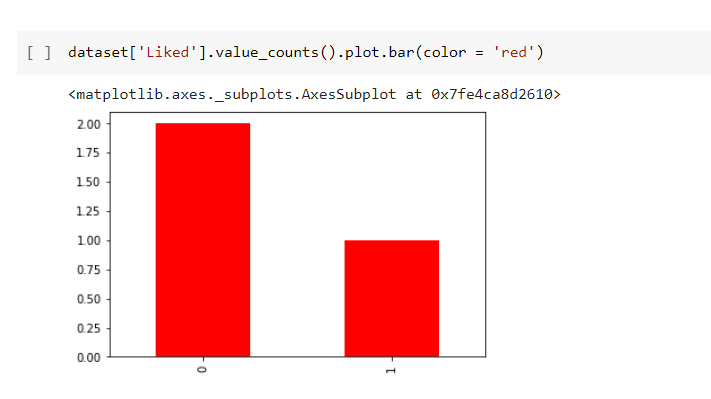




**Test run results:**

****

****

****

**CONCLUSION**

The proposed work starting from the analysis of different studies provided in the literature, provides a classification of sentiment classification approaches with respect to features/techniques and advantages limitations, tools for sentiment analysis with respect to the different techniques used for sentiment analysis.

The sentiment classification approaches can be classified in machine learning. The machine learning approach is used for predicting the review based on trained and test data sets.

Thus by our project we conclude that among the 6 applied Machine learning Models, LOGISTIC REGRESSION has the highest accuracy of 78%.

Then we have also tested the same models for the Given Test run instance by our faculty. Those also gave the same results.

**REFERENCE:**

[1] Mikel Joaristi, Edoardo Serra, Francesca Spezzano "Evaluating the Impact of Social Media in Detecting the Restaurants Violating the Health Norms" In 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

(ASONAM). [2] J. P. Schomberg, O. L. Haimson, G. R. Hayes, and H. Anton Culver. Supplementing public health inspection via social media.

PLOS ONE, 11(3), 03 2016..

[3] A. Sadilek. S. Brennan, H. Kautz, and V. Silenzio. nemesis: Which restaurants should you avoid today? In HCOMP, 2013. [4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. The Journal of

Machine Learning Research, 12:2825-2830, 2011. [5] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In ACL (System Demonstrations), pages 55 60, 2014.

[6] K. Lee, A. Agrawal, and A. N. Choudhary. Mining social media streams to improve public health allergy surveillance. In

ASONAM, pages 815-822, 2015. [7] K. Lee, A. Agrawal, and A. Choudhary. Real-time disease surveillance using twitter data: demonstration on flu and cancer.

In KDD, pages 1474-1477, 2013. [8] A. Lamb, M. J. Paul, and M. Dredze. Separating fact from fear.

Tracking flu infections on twitter. In HLT-NAACL, pages 789 795, 2013. [9] J. S. Kang, P. Kuznetsova, M. Luca, and Y. Choi. Where not to eat? improving public policy by predicting hygiene inspections

using online reviews. In EMNLP, pages 1443-1448, 2013.

[10] M. Dredze, M. J. Paul, S. Bergsma, and H. Tran. Carmen: A twitter geolocation system with applications to public health. In

AAAI/HIAI, pages 20-24, 2013.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer.

Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, pages 321-357, 2002. [12] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In EMNLP, pages 1568-1576, 2011.