ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

Social Media Content Classification and Detection for Trend Analysis Using Machine Learning Techniques

Aanal S. Raval, Ruchi K. Oza, Aakanksha V. Jain

M. E Students, Department of Computer, Silver Oak College of Engineering and Technology, Ahmedabad, Gujarat, India

ABSTRACT: To know trend in any field for business application or ant kind of survey, whether in entertainment, sports, academic, products, politics, etc., the best way is to consider social media data for peoples' taste and preferences. It is the better option as any user may or may not participate in the product survey.So here, we consider twitter data for this analysis. From this vast data, videos, links, audios are eliminated, only images text messages are considered. For text data, the words that doesn't matter i.e., background words are eliminated and words of interest are classified into categories. On images, logo detection is performed for the knowledge of product class. This categorical text classification and logo identification, enables to acquire the opinion of customers regarding a specific brand and displays concealed information, facts and awareness necessary for making decisions.

KEYWORDS: text preprocessing, embedding, classification techniques, anchor box representation, feature extraction

I. INTRODUCTION

Social media contains huge and voluminous data. Media like twitter and whatsapp contain information of users as well as their sentiments regarding a particular issue. Instead of feedbacks and comments posted on respective platform, social media can be used as source to know users taste. Here twitter data is considered. The idea for detection from messages is motivated by the fact that if a user is interested in issue, event, article etc, he or she will obviously show some frequent or noticeable behavior, such as replicate, reproduce or sharing tweets. Thus these freely available messages that contain sentiments can be analyzed for knowing business trend, irrespective of whether a user participates in specific product survey or not. From this data, background words are eliminated and words showing some event or issue i.e., words of interest are considered, and thus dataset is formed. To classify this enormous data, we considered bag of words model for searching similarity or dissimilarity between words. This classification is done in multi labels by assigning a label to each sample.

Secondly, it is possible that a user will never comment or replicate the tweets regarding some product, although he or she continues to prefer it over others. So preferences of these types of users can be detected by using images they post, checking in it whether it contains any logo or not. If the image contain logo, it can be detected and its product name can be obtained by feature extraction on that detected region. Thus it includes two stages- anchor box representation i.e., detection of region in the image and acknowledgement of logo with its legality towards a class i.e., feature extraction on detected region for retrieving product name.

II. RELATED WORK

In paper [5], logo detection consists of two stages, universal logo detector- predicts bounding boxes which are then classified by nearest neighbor search and few shot logo recognizer for feature extraction. For stage1, they used Faster RCNN, SSD, Yolov3. For stage2, SE-Resnet50 works as feature extractor with distance metric learning using proxies.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

In paper [1], For text classification, dataset is first preprocessed, then word2vec is used. Topics are generated by using LDA (Latent Dirichlet Allocation). This system is trained on Reddit data and is applied on twitter data.

III. METHOD FOR TEXT

A. Dataset and Preprocessing

For text classification, here we consider twitter dataset for testing. The whole mentioned methodology is trained on google news and tested on sample twitter dataset. For training purpose, fetch_20 Newsgroups is considered [1].

Preprocessing includes conversion to lower case, elimination of URLs, stop words, tokenizing, lemmatizing by WordNet Lemmatizer, and truncation by 5-fold cross validation.

Thereafter, Tf-idf scheme is applied, which refers to term frequency inverse document frequency. Term frequency is the original coarse count of term in document given by:

$$tf_{ij} = f / (\sum_i f_{ij})$$

Inverse document frequency is a measure of how much information the word provides, i.e., if it's common or rare across all documents and is given by:

$$Idf_I = \log (N / df_i)$$

So, tf-idf is:

$$tfidf(i,j,I)=tf(i,j).idf(I)$$

B. Word Embedding:

Next, Word2vev embedding is used to generate vector space of unique words. Word2vec is network with two layers which is trained to generate linguistic context of words. Here the context is set of multiple words. As the dataset is large, common bag of words model is considered. So this outputs the probability of certain words within a specified context.

C. Classification:

For classification stage, here we tested three techniques:

First we tested with eXtreme Gradient Boosting [2], as this algorithm is more efficient in text classification than any other currently present for handling sparsity patterns. Unlike random forests, XGBoost provides methodical and systematic solution and aggregates prediction power of multiple learners. So here the final output is not only the majority vote of predictors but the averaged output from all. Thus it provide parallelized implementation approach by building trees sequentially.

Second test we conducted with CatBoost [4], resting on decision trees which are gradient boosted. At the training time, this algorithm builds a set of decision trees consecutively, with each tree having reduced loss as compared to previous.

Third is Light GBM [3], which grows trees vertically that means it grows leaf wise instead of level wise to reduce loss. So, it will select the leaf with max delta loss.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

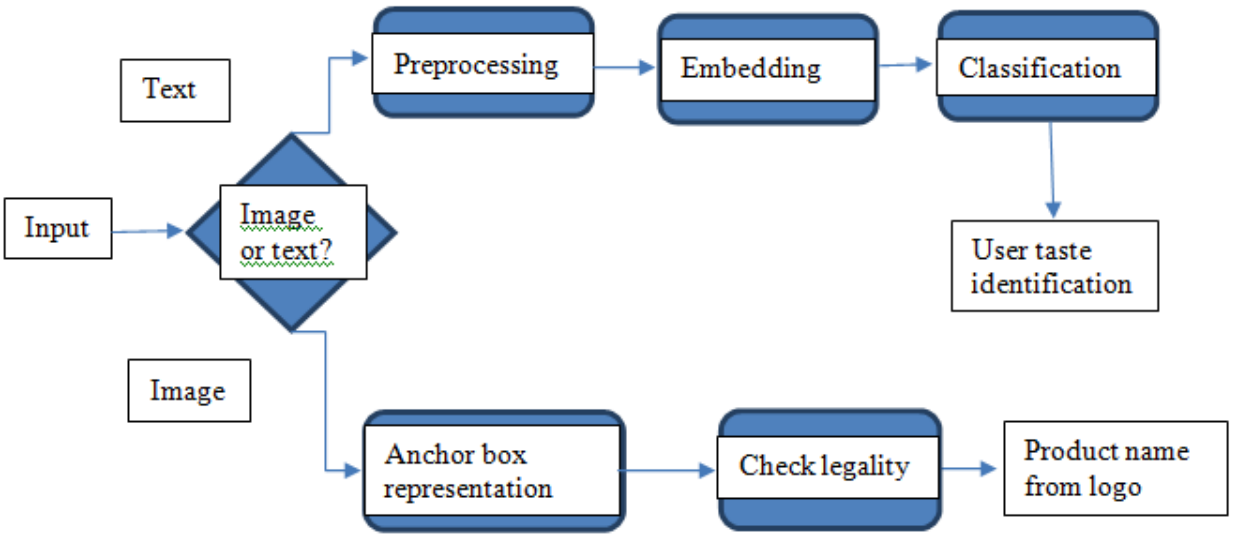

Figure-1

IV. METHOD FOR IMAGES

A. Dataset:

We created a small sample dataset with 17 different logo classes each with 10 images for both training and testing.

B. Methodology

Identification of logo includes two stages: first is anchor box representation where bounding box are predicted along with location of logo in the image. Secondly, logo acknowledgement and legality where features are extracted and the class of logo is predicted [5].

For both the stages, we implemented and tested different combinations of algorithms for accuracy. First test is use of CNN [7] for both the stages. It is the simplest structure where combination of convolution with RELU function and pooling work for feature learning and fully connected layers with softmax function does the work of classification.

Second, SSD (single shot detection) with Mobilenet for bounding box prediction and SE-RESNET101 (residual network with squeeze and excitation block) as feature 5extractor for logo product class prediction.

Third test we made on LeNet [6] for both the stages, which consists of two sets of convolutional and average pooling layers, followed by a flattening convolutional layer, then two fully-connected layers and finally a softmax classifier.

Fourth combination is of YOLO (you only look once) for bounding box prediction and Inception v3 for recognition [20].

Figure-1 shows the flow of methodology.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

V. RESULTS

A. *Text Classification*

Table-1 shows the results of CatBoost and XGBoost compared with other methodologies. XGBoost has quite good result of 74% accuracy than all of those mentioned. Light GBM shows accuracy of 77.32%, which is better than XGBoost. Besides CatBoost outperforms all with accuracy of 85%.

Light GBM and XGBoost are faster, i.e., took less time to learn while CatBoost is accurate but on this dataset it took 12 hours to learn. Thus, CatBoost is much slower.

B. *Logo Detection:*

CNN performed well (99%) on 17 logos, but as this was tested with new class its accuracy also decreased. LeNet also performs well with 99% accuracy, with 15 epochs. Combination of SSD-mobilenet with SE-Resnet101 yield accuracy of only 84.39%, but it is faster as well as performed well for smaller dimensions of logos in image. Yolo with Inception v3 achieved 77% on this dataset.

Model name	Accuracy	Precision	Recall	F1score
SVM	67	68	67	67
MNB	66	67	66	66
Decision tree	55	56	55	55
Random forests	56	58	56	56
Rocchio classification	69	74	69	70
XG Boost	74.24	76	74	75
Light GBM	77.32	78	77	77
Cat Boost	85	85	85	85

Table-1

Model name	Accuracy (for these 17 logos)
CNN	99
LeNet	99.85
SSD-mobilenet and SE-Resnet101	84.39
YOLO-Inceptionv3	77

Table-2

V. CONCLUSION

For text classification, results convey that Light GBM can be used where time period matters. But CatBoost once trained will display more accurate results. For logo, SSD-mobilenet with SE-Resnet101 and LeNet both are good options.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

The methodology is able to hypothesize user preferences from twitter data in an effectual and productive manner, by training on fetch_20 Newsgroups. So these data can further be used in economical surveys, marketing as well as tracking social, political, educational, medicinal issues. For literature, labels can be specialized as fiction, comics, novels, academy, science books etc.

Further work includes consideration of data from other social media. Besides, links and videos also can be considered in dataset. Apart from these, simply text summarization with NLP can be performed on links to documents, which can also revert field of user interest. If the dataset is mixture of images and text, a combination of UMAP (Uniform Manifold Approximation and Projection) and InceptionResnetv2 can be used to categorize(if need is there), isolate and recognize images for further mentioned processing.

REFERENCES

[1] Angel Fiallos and Karina Jimenes, "*Using Reddit Data for Multi-label Text Classification of Twitter Users Interests*", 2019 IEEE.
[2] Tianqi Chen and Carlos Guestrin, "*XGBoost: A Scalable Tree Boosting System",* KDD'16, August 13-17, 2016, San Francisco, CA, USA.
[3] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma1 , Qiwei Ye, Tie-Yan Liu, "*LightGBM: A Highly Efficient Gradient Boosting Decision Tree",* 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
[4] LiudmilaProkhorenkova, GlebGusev, AleksandrVorobev, Anna VeronikaDorogush, AndreyGulin, "*CatBoost: unbiased boosting with categorical feature",* 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.
[5] Istv´anFeh´erv´ari and SrikarAppalaraju, "*Scalable Logo Recognition using Proxies",* 2019 IEEE Winter Conference on Applications of Computer Vision.
[6] YannLeCun, Leon Bottou, YoshuaBangio and Patrick Haffner, "*Gradient Based Learning Applied To Document Recognition",* PROC OF THE IEEE, NOVEMBER 1998.
[7] Simone Bianco, Marco Buzzelli, Davide Mazzini, and RaimondoSchettini, "*Logo Recognition Using CNN Features",* ICIAP 2015.
[8] R Nivedha and N. Sairam, "*A Machine Learning based Classification for Social Media Messages",* School of Computing, SASTRA University, Thanjavur – 613 401, Tamil Nadu, India.
[9] Hari Krishna Kanagala, Dr. V. V. Jaya Rama Krishnaiah, "*A COMPARATIVE STUDY OF K-MEANS, DBSCAN AND OPTICS",* 2016 International Conference on Computer Communication and Informatics (ICCCI -2016), Jan. 07 – 09, 2016, Coimbatore, INDIA.
[10] Alrence Santiago Halibas, AbubuckerSamsudeenShaffi and Mohamed Abdul Kader Varusai Mohamed, "*Application of Text Classification and Clustering of Twitter Data for Business Analytics",* 2018 IEEE.
[11] Daniel Mas Montserrat, Qian Lin, Jan P. Allebach and Edward J. Delp, "*Scalable Logo Detection and Recognition with Minimal Labeling",* 2018 IEEE Conference on Multimedia Information Processing and Retrieval.
[12] Matthew Long, Jiao Yu, and AnshulKundani, "*Twitter Classification into the Amazon Browse Node Hierarchy",* December 12, 2014.
[13] Jie Hu, Li Shen and Gang Sun, "*Squeeze-and-ExcitationNetworks",* 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.
[14] Kowsari, Kamran and Brown, Donald E and Heidarysafa, Mojtaba and JafariMeimandi, Kiana and and Gerber, Matthew S and Barnes and Laura E, "*HDLTex:Hierarchical Deep Learning for TextClassification",* 2017, IEEE.
[15] NithirojTripatarasit, "*Logo Detection Using PyTorch",*PyCon Thailand 2018.
[16] Nabin Sharma, RanjuMandal, Rabi Sharma, Umapada Pal ,Michael Blumenstein, "*Signature and Logo Detection using Deep CNN for Document Image Retrieval",* 2018, IEEE, 16th International Conference on Frontiers in Handwriting Recognition (ICFHR).
[17] Kai Huang, Rishita Roy and Justin Yu, "*LogoDetector: Logo Recognition using YOLO".*
[18] AndrasTuzk, Christian Herrmann, Daniel Manger and JurgenBeyerer, "*Open Set Logo Detection andRetrieval",* In Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 5: VISAPP, pages 284-292.
[19] Hang Su, Shaogang Gong and Xiatian Zhu, "*WebLogo-2M: Scalable Logo Detection by Deep Learning from the Web",* CVF, 2018.
[20] Angelo Monteux, "*LogoHunter: Brand Detection As A Service*", Machine learning project developed at Insight Data Science, 2019 AI session. (From github)