International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

A Survey on Hashtag Generation and Prediction for Images and Text

Nischitha N [1], Shashank Bharadwaj R [2], Suchethana Swaroopa PN [3], Sudeep S [4],
Dr. Ravi Kumar V [5], Tanuja Kayarga [6],
[1,2,3,4] UG Students, Department of CS&E, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India
[5] Professor and Head, 6 Assistant Professor, Department of CS&E, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India

***Abstract* - Social media networking has changed the way a business is functioned. Hashtags help one to promote and set a brand for oneself. Hashtag is a string prefixed with a hashtag in the very beginning (#). On social media platforms, content can be traced using hashtags, therefore making hashtags like a key-pair value in a dictionary. Machine learning is a method of finding patterns in the data, and using it, we tend to find and predict relevant hashtags for images and texts in order to have better engagement on social media platforms. In this paper we are exploring the possible better solutions through various approaches.**

Keywords - Hashtag, Machine learning, Convolutional Neural Network, Computer Vision, Natural Language Processing

INTRODUCTION:

Hashtags are used almost on each and every post that is made on social media platforms for a better reach. Approximately, 90 lakh photos, videos and reels are shared on instagram everyday and on an average, it is estimated that each post uses around 5 tags. On twitter, hashtags are used to make a topic trend that makes an impact, either positive or negative. In this way, hashtags are used to gauge the stats and topics on the net.With the help of ML, we accurately suggest hashtags for images and texts, which therefore when used, gives the user a better reach on social media. Finding the right hashtags for your post is an important task, as not doing so, will not give the traffic required for the business on social media platforms. There are dedicated hashtag analytics websites as well in order to provide the best hashtag that gives the maximum reach. The first ever hashtag was used in 2007 on a social media post, and ever since the field of branding has changed.

LITERATURE SURVEY:

In [1] the authors present a dataset called Harrison dataset. It is created with the help of a website called top-hashtags.com (https://top-hashtags.com/instagram), which is a hashtag ranking website used by many people to get hashtags based on a particular topic. After a bit of data cleaning, that is, removal of repeated images, around 91,000 images are obtained with their respective hashtags, with an average of 15.5 hashtags associated with each image. Later, hashtags are cleaned, like removal of hashtags associated with other languages. After this, lemmatization is applied to all the hashtags. It is a process of grouping all words occurring from the same root word (talk, talking, talks etc.). This creates 1,65,000 unique hashtags. Top 5000 hashtags constitute about 75% of the hashtags. In this manner, the dataset consists of around 1000 classes. VGC-16 feature extractor is used and 52.2% accuracy has been obtained.

In [2] the authors propose a way to transfer the knowledge learnt from one source to another, and hence says that the model and the training dataset need not be in the same feature. The authors take an example of web-document classification and explains, stating that after the model is trained, the paper or document is classified into several predefined categories. But for a classification on a new dataset or a new document, there might be a deficiency of labelled knowledge. In this case, the data has to be trained once again, making it a heinous task. Therefore with the help of knowledge transfer or transfer learning, one can take the pretrained models and classify the newly available data without necessarily training it each and every time an upgrade is made. However, this paper mainly focuses on knowledge transfer in classification, clustering and regression. In this paper the methods used are SVM with accuracy 93% SGD with accuracy 84% and TrAdaBoost with accuracy 92%.

In [3] the authors put forward about zero shot learning, which is the process of classifying the images at the very first glance, that is, which is not in the trained data. ResNet feature extractor has been used. Here, the authors inspect sevelar zero shot classifying techniques in depth, for various datasets, from small to large-scale. Evaluation is made based on three factors, that is, methods, datasets and evaluation protocols.

In [4] the authors propose the concept of how product based companies use microblogging websites like twitter to assess their products and collect the user input based on the reviews they provide. The work clusters the overall reviews into 3 categories, positive, negative and neutral. They also measure the performance of 3 models, namely, “a unigram model, a feature based model and tree-kernel based model” with attaining the accuracy of 75.39%. On performing the above mentioned task, it is observed that the tree-kernel based model outrange the other two models by a huge margin. Later, a combination of two models is performed,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

which again outperforms the individual models. The work also suggests that hashtags and emoticons add value to the classifier, but only fractionally.

In [5] the author proposes the concept of "image captioning" which is to categorize the parts of images to multiple categories, for example, people, animals, tables etc. This is quite interesting as it deals with both NLP and CV simultaneously. On testing the dataset, 108 captions were generated out of which 55 were positive and the remaining, negative. Positive tends towards results that were related to the image, while the negative one was completely irrelevant.

In [6] The authors take inspiration from journalism about the 5 W's used in it. Based on this, the 5WTAG algorithm is formed.
The 5 W's are:
Who
What
When
Where
hoW
It further proves how events on social media are event-centric. It also stresses on the inverted pyramid principle used in journalism. pLSA,LDA feature extractor has been used attaining 64.8% accuracy.

In [7] the work uses data mining to find patterns that can be employed in real world apps to procure knowledge. Data mining is used to derive meaning patterns therefore helping the users to find relationships and correlations in databases. It also introduces a framework called PMHRec, short for Pattern Mining for Hashtag Recommendation.

In [8] this work, the authors stress how hashtags can be used by companies to promote their business with the help of hashtags in the near future. They also compare multiple models and suggest the best one which gives the highest accuracy. This work uses Harrison dataset and algorithms such as zero-shot classifiers. To test on real world photos, a tool called instaloader is used, which downloads publicly displayed instagram images. The feature extractor used is VGG-19 with accuracy of 68.14%.

In [9] NLTK is introduced. The Natural Language Toolkit (NLTK) is a collection of program units, data sets etc. NLTK is written in Python and spread under the General Public License. In the proposed system, Naïve Bayes classifier permits to categorize the hashtags grounded on the examination given by the explicit algorithms that trained using large training data-sets.

Presently, there are scarce hashtag generating applications. Hence the user has to physically explore for hashtags. Existing applications have hashtags which are not updated because of which one cannot find out if the hashtag is trending or not, hence the expected reach is lowered. The disadvantage is that the current applications are unable to assure the genuineness of the seller uploaded content. Seek metrics, All-hashtags are few examples of aforesaid applications which can produce hashtags only for input of small size. When the input is heavy, the performance's accuracy drops.

Based on this paper, Hashtag Generator and Content Authenticator is an application that lets the users discover hashtags which are popular to get a desirable amount of reach from the audience. The Authors have used a Sequential approach to get the output.Feature extractor is done based on the scale invariant feature transformation.

In [10] the paper proposes to find mislabeled data. Mislabeled samples are hard to avoid while constructing large datasets. A small number of mislabeled examples are considered and reviewed by this approach.

Non experts construct Large scale datasets which are much prone to errors. To minimize such errors Heuristic approach is used. In this paper Authors have considered an example of building ImageNet, it is stated that the ImageNet dataset has only 0.3% label noise errors across all synsets. The goal is finding a method that has fewer human interventions and also utilizes less energy to label large datasets.

GoogleNet convolutional neural network model was used as a Feature Extractor. Experiments with three datasets were conducted: ImageNet, UCI character recognition and MNIST digit recognition. The level of label noise found in ImageNet Dataset is equivalent to the stated value of 0.3%. The outcomes displayed that the method used was able to remove most of the mislabeled examples from the MNIST digit and UCI character recognition datasets.

This paper presents an approach to find mislabeled samples in large datasets. In 18 image classes, 92 mislabeled samples were found. The approach presented needs 9 times fewer samples to find the same number of mislabeled samples.

In [11] the work proposes a method to remove mislabeled data. Mislabeled examples in the training data disturbs the learning process and have an opposing consequence on the performance of supervised classifiers. In this paper Mistrustful samples are considered for inspection and an approach to remove mislabeled samples is presented. Results of the Experiments conducted on

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

the datasets shows that support vector machines (SVM) are capable of capturing 85% to 99% of label noise samples. A new method is proposed that repeatedly builds two-class SVM classifiers based on the training dataset's non-support vector samples, later it is followed by an expert confirming by-hand the support vectors to identify mislabeled samples based on the classification score. The method discussed is parameter independent which was proved by the experiments conducted on four datasets. The results proved to be advantageous for building labelled datasets as most of the noise can be removed by re-examining the labels of the support vectors.

It is experimentally shown that label noise samples are chosen as outliers and the support vectors of the one-class support vector machines (OCSVM) and two-class support vector machines (TCSVM). The Results show that the performance of TCSVM is greater than OCSVM in choosing support vectors from the label noise samples. TCSVM also removes a greater number of label noise samples and at the same time reviews a smaller number of samples than OCSVM.

In [12] the work brings forward about the formation of the ImageNet dataset. The ImageNet Large Scale Visual Recognition Challenge is a standard in classification and discovery of many object groups and masses of images. The paper aims at describing the formation of a standard dataset. Following points are discussed:

- The challenges faced while collecting large scale annotations
- High spot important breakthroughs in object recognition
- A thorough breakdown of the current state of the field image classification and object detection
- Comparison of the state-of-the-art computer vision accuracy with human accuracy.

This paper describes the collection of large-scale data processes of **ImageNet Large Scale Visual Recognition Challenge** (ILSVRC). Most effective algorithms are considered and summarized. Also, the success and failure modes of the algorithms considered are analyzed.

In [13] offers image security and digital watermarking techniques. The number of images uploaded into social media is increasing rapidly everyday because of the increasing number of people using social media. Taking necessary precautions before uploading images on social media is required because of the current issues such as identity fraud, tarnished metadata from the images. For this watermarking is helpful. Selected metadata of the image is used for enhancing the security of digital watermarking processes. In this paper, for experimentation purposes an android application called MyImage is used to instrument both invisible and visible watermarking algorithms. The images which were watermarked were used on four different social media platforms and the metadata for the watermarking process were analyzed.

In this process a translucent layer is created which overlaps the image. This technique helps to protect the image from being copied and also maintains its originality. This paper also discusses the invisible watermarking technique that is the frequency-domain technique. In this technique the whole image is transformed into frequency domain coefficients and then the image is rooted into these coefficients. This paper aims at discussing the usage of metadata to protect the images uploaded into social media.

COMPARISON AND ANALYSIS OF DIFFERENT PAPERS

Sl.no	Objectives	Methodology/ Models Used	Result/ Outcomes	Dataset	Drawbacks
1	Recognising objects at the very first look and trying to generate hashtags for real-world objects. For the evaluation of the photos, a bottomline framework is used, which consists of visual feature extraction based on CNN.	CNN, VGG-Object, VGG-Scene	In this work, on applying two visual features, object-based features and scene-based features, the baseline models were evaluated.	Harrison	With only a few images, not too many categories can be seen, which will not be very beneficial in real-world applications.
2	This work not only tells how training and newer data need not be in the same space, but also talks about progress of TL in classification, regression etc.	Transfer learning, ML, Data mining	This work suggests "what to transfer" and "when to transfer". Also, it proposes where improvement needs to be made, like negative transfer, different feature spaces etc		Transfer learning is not applied on all of the ML algorithmic models. Only three models are taken into consideration here.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

3	This work challenges the n number of proposed systems for zero shot learning. It proposes a new threshold factor by unifying rules and data splits.	Attribute Label, Deep Visual Semantic and Structured Joint Embedding	In this work, several algorithms and models have been applied to various datasets with different size and features. Therefore, it concludes the pros and cons of zero-shot learning.	Attribute Pascal and Yahoo Animals with Attributes Caltech-UCSD-Birds	Results obtained in a generalized zero-shot classification is significantly lower than that performed on training data.
4	This work inspects the sentiments on twitter data and analyses it. In addition to it, it also contributes to POS, tree kernel to obviate the need for tedious feature engineering.	Prior Polarity Scoring, Dictionary of Affect in Language, Partial Tree, WordNet	This work adds an accuracy of 4% to the previously advanced unigram model.	Commercial source	Emoticons with prefixed meaning were used, which is not related to the text, therefore, it doesn't predict satire and sarcasm.
5	This work uses Harrison dataset to first generate a hashtag for a given image and upon that, give a small caption/anecdo-te based upon the hashtag generated.	Image caption, Computer Vision, Natural Language Processing, RNN, CNN-LSTM,	This work proved to be accurate. To test the accuracy, random people were taken from the internet to test the model as well and the results thus observed were satisfactory.	Harrison	Although the author proposes that a proper caption is generated, we see that the captions would not make a proper sense most of the time.
6	This in work, they retrieve information about the tweet and have created an algorithm named 5WTAG, which properly extracts information from microblogs.	probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA), 5WTAG	This work introduces us to an algorithm called 5WTAG. Taking example from real life journalism of 5W's (What, Where, Why, When and hoW), it detects real-life incidents through tweets.	Sina Weibo	When hashtags that are similar are cluttered, the model tends to lose its accuracy. And this becomes a problem when too many people tweet from the same spot.
7	This work introduces a framework known as PM-Hrec (Pattern mining for Hashtag Recommendati-on). Here, the tweets are downloaded and their high average patterns are searched for.	PM-Hrec, UPM, FPM, WFPM, SPM	This paper movies the top utility average pattern of the tweets that are downloaded. The disadvantage is that it consumes a lot of space.	Twitter API to collect and store the tweets	This work consumes too much of memory than other similar models for the same problem statement.
8	In this paper, generating hashtags are introduced from images online. Here, ML is used to predict hashtags for realistic images and makes choices and branding much easier.	Zero-shot classification, Supervised ML algorithms, ImageNet classification, DCNN	This work has an accuracy of about 85%. It generates multiple hashtags for a single image. These hashtags can be directly used in social media accounts like Instagram, Facebook, Twitter etc. to get a better engagement on the respective platforms.	Harrison	Very few categories have been classified in the Harrison dataset, and on using the same, this work does the same. Also, some images have a few hashtags for a number of images (around 2).
9	Hashtag Generator and Content Authenticator is an application that lets the users discover hashtags which are popular to get a desirable amount of	Analysis and Requirement Gathering, Image Authentication using Metadata, Image Feature Extraction, Text Analysis, Recommending Hashtags, Classification	Images undergo an image authentication phase followed by the display of tags which are forwarded to the image analysis component and relevant keywords are generated. Keywords are	Not mentioned	Cannot be accessed by 3rd party users. Image authentication is only for a small range of domains such as fashion, nature and travel.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

	reach.	and Filtering Hashtags	analyzed and relevant hashtags are suggested.		
10	An approach to select a small number of mislabeled examples which will be reviewed by experts to find some mislabeled examples in a large scale dataset (ImageNet).	Support Vector Machines, ImageNet Dataset, Character Recognition Dataset,	This paper presents an approach to find mislabeled samples in large datasets. In 18 image classes, 92 mislabeled samples were found. The approach presented needs 9 times lesser examples to find the same number of mislabeled samples.	ImageNet dataset	No experiments were conducted on other different features and algorithms. The class pairs chosen must be confusing in order to find label noise.
11	Mistrustful samples are considered for inspection and an approach to remove mislabeled samples is presented.	Support Vector Machines, One-class Support Vector Machines (OCSVM), Two-class Support Vector Machines(TCSVM)	The performance of TCSVM is greater than OCSVM in choosing support vectors from the label noise samples. TCSVM also removes a greater number of label noise samples and at the same time reviews a smaller number of samples than OCSVM.	UCI letter recognition. MNSIT digit dataset	This approach is not feasible to find label noise in multi-class problems.
12	The ImageNet Large Scale Visual Recognition Challenge is a standard in classification and discovery of many object groups and masses of images. This paper aims at describing the formation of a standard dataset.	ILSVRC dataset	This paper has an accuracy of 94.6% and describes the collection of large-scale data processes of ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Most effective algorithms are considered and summarized. Also, the success and failure modes of the algorithms considered are analyzed.	ILSVRC dataset	There are many paths of improvement for ILSVRC. As datasets grow, it becomes dreadful to fully annotate them.
13	To solve current issues faced such as identity fraud, tarnished metadata of images uploaded into social media by using digital watermark techniques.	MyImage, a java based android application, visible watermarking technique, invisible watermarking techniques (frequency-domain technique).	Images were uploaded into four different social media platforms. They were downloaded back to test the presence of watermarks. All images watermarked using visible watermarking techniques passed the tests.	Not mentioned	Visible watermarking technique, is prone to cropping attacks. For Invisible watermark, the metadata remained untarnished only in Twitter.

Table 1: Summary of survey papers

CONCLUSION

Hashtag has changed the way of sharing information. It also promotes business growth. The previous work had a smaller dataset with fewer categories while summarising the survey papers. A method is proposed where hashtags can be generated based on text and images with better accuracy considering larger datasets. The proposed method uses Computer Vision and Natural Language Processing to recommend hashtags, which will help numerous social media handles to benefit out of it, which we plan to improve in our work using a different dataset.

ACKNOWLEDGEMENT

We express our gratitude towards the guidance provided by our guide Dr. Ravi Kumar V, Professor and Head, Department of Computer Science and Engineering, Vidyavardhaka College of Engineering.
We sincerely thank our guide Professor Tanuja Kayarga for their encouragement and guidance in carrying out this work.
We also thank our mentors and faculty members who guided us throughout the research.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

REFERENCES

[1] Minseok Park, Hanxiang Li and Junmo Kim “Harrison: A benchmark on HAshtag Recommendation for Real-worl Images in SOcial Networks” A Benchmark on Hashtag Recommendation for Real-world Images in Social Networks, arXiv:1605.05054v1 [cs.CV] 17 May 2016.
[2] Sinno Jialin Pan and Qiang Yang “A survey on transfer learning” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 10, OCTOBER 2010.
[3] Yongqin Xian, Bernt Schiele and Zeynep Akata “Zero-Shot Learning - The Good, the Bad and the Ugly” 2017 IEEE Conference on Computer Vision and Pattern Recognition.
[4] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau “Sentiment Analysis of Twitter Data” Department of Computer Science Columbia University New York, NY 10027 USA.
[5] Shivam Gaur “Generation of a short narrative caption for an image using the suggested hashtag” 2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW).
[6] Zhibin Zhao, Jiahong Sun, Lan Yao, Xun Wang, Jiahong Chu, Huan Liu, and Ge Yu “Modeling Chinese Microblogs with Five Ws for Topic Hashtags Extraction” TSINGHUA SCIENCE AND TECHNOLOGY ISSNll1007-0214ll02/09llpp135–148 Volume 22, Number 2, April 2017.
[7] Asma Belhadi, Youcef Djenouri, Jerry Chun-Wei Lin andAlberto Cano “A Data-Driven Approach for Twitter Hashtag Recommendation” Dept. of Computer Science, USTHB, Algiers, Algeria.
[8] Prajwal Bharadwaj N, Taurunika Shivashankara, Madhushree S, Prajwal K, Sachin D N “Prediction of Hashtags for Images” International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 9 Issue 07, July-2020.
[9] Kavinga Yapa Abeywardana, Ginige A.R., Herath N., Somarathne H.P., Thennakoon T.M.N.S." Hashtag Generator and Content Authenticator"(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 9, 2018
[10] Rajmadhan Ekambaram, Dmitry B. Goldgof, Lawrence O. Hall "Finding Label Noise 332Examples in Large Scale Datasets" 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) Banff Center, Banff, Canada, October 5-8, 2017
[11] Rajmadhan Ekambaram, R., Fefilatyev, S., Shreve, M., Kramer, K., Hall, L.O., Goldgof, D.B. and Kasturi, R.: Active cleaning of label noise. Pattern Recognition, 51, pp.463-480. (2016)
[12] Olga Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), pp.211-252 (2015)
[13] M. bin Jeffry and H. Mammi, "A study on image security in social media using digital watermarking with metadata - IEEE Conference Publication", Ieeexplore.ieee.org, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8270435/.