Hashtag Generator and Content Authenticator

Kavinga Yapa Abeywardana[1], Ginige A.R.[2], Herath N.[3], Somarathne H.P.[4], Thennakoon T.M.N.S.[5]
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka

Abstract—**In the recent past, Online Marketing applications have been a focus of research. But still there are enormous challenges on the accuracy and authenticity of the content posted through social media. And if the social media business platforms are considered, majority of the users who try to add a market value to their own product face the problem of not getting enough attention from their target audience. The purpose of this research is to develop a safe and efficient trending hashtag generating application solution for social media business users which generates trending and relevant hashtags for user content in order to get a broad reach of target audience, automatically generates a meaningful caption to their relevant posts and guarantees the authenticity of the product at the same time. The user content is analyzed and filters the important keywords, generates a meaningful caption, suggest related trending keywords and generates trending hashtags to get the required reach for online marketers. Additionally, the marketing products' content authentication is ensured. The application uses Natural Language Processing, Machine Learning, API technologies, Java and Python technologies. A unique database is assigned to users which contains rankings for each user. The target audience who engages in buying products get to know about the status of the sellers with respect to authenticity of the content. It is believed that the application provides a promising solution to existing audience reach problems of online marketers and buyers. The significance of this system is to help marketers and buyers to engage in online buying and selling with much effective, reliable and safer ways. This mitigate the vulnerability of bad social media marketing influences and helps to establish a safe and reliable online marketing practice to make both sellers and buyers happy. This paper provides a brief description on how to perform an organized online marketing discipline via the Trending Hashtag Generator & Image Authenticator application.**

Keywords—Hashtags; social media; NLP; machine learning; REST API; content authentication

I. Introduction

A hashtag is a name or an identifier that resolves to a description of its referent. In other words hashtag is a keyword or phrase preceded by the hash symbol, written within a post or comment to highlight it and facilitate a search for it. In the present day, hashtags are immensely used for brand promotion and social media discussions. In principle hashtags facilitate powerful identification functionality to any kind of HTTP based services (social media platforms etc.). Essentially, by including hash marks in your post, provided the appropriate privacy settings are in place it can be indexed by the social networks so that it is discoverable by everyone, even if the user is not subscribed to your account updates. There are countless number of individuals and merchants who intend to sell a genuine product by identifying and focusing on a targeted audience. Since hashtags play a vital role in online marketing, aforementioned people try to find the most relevant hashtags to increase the audience reach for their products. However most of them end up having a hard time to identify the most relevant and trending hashtags at particular given time. The major problem faced by the user is "unavailability of a platform that is made for clients which can analyze the post (photo/text content) and fetch trending hashtags specific for the relevant post, auto generate a meaningful caption, categorize the hashtags according to the popularity which can get a considerable awareness from target audience in a completely new and user-friendly way". Even though the content is published along with some hashtags manually, the content uploader does not have any prior knowledge whether the hashtags he/she manually searched are trending or not at the current time. On the other hand, if an uploader uploads a digitally captured image of the product, the audience does not have any proof whether the content is genuine or not. It is a commonly known fact that buyers are reluctant to purchase a product when they have a suspicion about the authenticity of the advertised content. Considering these facts the authors have introduced the "Hashtag Generator and Content Authenticator" as a web application which lets users easily upload their content, get a categorized trending hashtags list along with a caption, select the preferred hashtags according to his/her preferences and push the content to the relevant social media platform. The web app lets the consumer to generate the desired amount of reach via hashtags which could gain much required attention for commercial and noncommercial purposes. The authors have developed the app to authenticate the uploaded content with a specialized tag for the customers to identify pre authenticated content. This will prevent or hinder unauthorized copying and editing of the content which ensures authenticity. It will help uploaders to prove the authenticity of their content to their audience which is crucial from sales perspective. These processes are done using API technologies, Natural Language Processing (NLP), machine learning and data forensic technologies. The proposed features are unique compared to existing tools due to the unavailability of all these features in modern social medial platforms and related applications. The authors have provided an accurate, user friendly web application interface to ensure the authenticity of the products to be sold, making things easy for both the sellers and customers (target audience).

II. Literature Review

A. Analysis of the usage of Metadata

The literature conferred by Hazinah Kutty Mammi and Mohd Aliff Faiz bin Jeffry[1] explore the practicability of securing the images by using metadata with digital watermarking. They have proposed to embed selected metadata to the image in a watermarking procedure. These embedded images were used in multiple social media platforms for analyzing the accomplishment of the project and proposed watermarking techniques. The literature discusses how metadata can be used to prove the originality of an image.

During the literature review, the priority was given to more modern approaches to establish a solution to prohibit users publishing downloaded images from the internet as their own .The research done by Walid Hussein, Osman Ibrahim and Mostafa A. Salama [2] on image processing using the signature verification techniques reviews how signature verification techniques can be done using an image.

B. Analysis of Web Crawling Algorithms

In the process of the literature survey, it was identified that the main web crawler approaches that can be used for comparison are based on the following characteristics.

- Resemblance and the relativity of the data that the crawler acquire
- Effective relevance forecasting to facilitate downloading the composition.
- Overall effectiveness of the crawler

In conclusion it was identified that the most beneficial algorithm presented is the ‘Focused Crawling’ algorithm due to its minimum response time. The research done by Handoko, et al. [3] on optimization of the focused crawler by using Genetic Algorithm has combined the focused crawler with genetic algorithm to resolve and fine tune web searching. It discusses about the complications caused by the local searching algorithms. Table 1 shows the comparison carried out on both approaches. The features used in the Focused crawler and the Genetic Algorithm will be used to achieve the goals in the proposed solution.

TABLE I. PRECISION OF THE WEB CRAWLERS’ COMPARISON DONE IN THE RESEARCH[3]

Category	BFS crawler’s precision	GA crawler’s precision
Education	90%	97%
Computer	85%	97%
Digital	80%	82%
Analog	63%	93%
Sport	90%	95%

Researchers have studied and developed multiple NLP related solutions over time to solve problems in various domains. They can be categorized as follows.

- Lexical and morphological analysis, noun phrase generation, word segmentation, etc.
- Semantic and discourse analysis, word meaning and knowledge representation.
- Knowledge-based approaches and tools for NLP [4].

Noun phrasing is considered to be an important NLP technique used in information retrieval. One of the major goals of noun phrasing research is to investigate the possibility of combining traditional keyword and syntactic approaches with semantic approaches to text processing in order to improve the quality of information retrieval [4].

The growing technology of NLP suggests that there are two possible scenarios for the future interactions between computers and humans: in the user-friendliness scenario, computers become smart enough to communicate in natural language, and in the computer friendliness scenario humans adapt their practices in order to communicate with, and make use of, computers [4].

Topic modeling is a type of statistical modeling for discovering topics that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

Recently, a number of studies have shown that the use of machine learning and text mining methods to automatically identify relevant studies has the potential to drastically decrease the workload [5].

Topic analysis is currently gaining popularity in both machine learning and text mining applications.

Automatic text classification for systematic reviews has been investigated by Bekhuis [6] who focused on using supervised machine learning to assist with the screening phase. Octaviano [6] combined two different features, i.e. content and citation relationship between the studies, to automate the selection phase as much as possible.

From a topic modelling perspective, Miwa firstly used LDA to automatically suggest topics for related keywords and reduce the difficulty of systematic reviews using an active learning strategy [6].

Since the emergence of topic models, researchers have introduced this approach into the fields of biological and medical document mining [7]. Such experiments proved LDA could be successfully applied to text classification. In the present day, LDA modeling is being developed for machine based communication purposes [7].

The trending hashtag recommendation problem addresses suggesting hashtags to explicitly tag a post made on a given social media platform, based upon the content and the context of the post. The issue of trending hashtag recommendation has emerged as a mainstream area of research overtime. “Hashtag recommendation for micro-blogs using topic specific translation” by *Ding, Q. Zhang, and X. Huang* [8][9] are researches based on NLP & Probability based algorithms by key phrase extraction and model them into topic specific translation. “Recommending #tags in twitter” by *E. Zangerle, W. Gassler, and G. Specht* [10] is also a research that targets microblogs but this is based on extracted hashtags ranking model. All these research efforts are specifically targeted at microblogs which is a highly specific area of content.

In the research “Semantic embedding from hashtags” by *Weston, S. Chopra, and K. Adams* [11] they used an NLP and ML based Convolutional Neural Network for hashtag recommendation with supervised word embedding. Comparatively this can be identified as a successful approach.

In present most of the research work utilize the advancements of ML to achieve their objectives. “User Conditional Hashtag prediction for Images” by *E. Denton, et al.* [12] is an approach that used ML along with the hashtags & contextual information about the user to perform hashtag prediction for user given image. Simply how user meta-data combined with images derived from a CNN can be used to predict hashtags. With the data, the researchers developed a user model which could be applied for a large dataset that is taken from Facebook. The user model primarily predicted hashtags, but the predicted hashtags were not “trending hashtags”. In this approach a hashtag embedding model will be used that trains with the collected data. This method would be very practical because of the availability of data.

Data extraction from social media platforms comes under the categorization of social media data mining. Mainly there are three different ways to harvest data from social media platforms. Those are through APIs, personal archives and scraping. Since most of the social media platforms have updated their restrictions on data extraction due to various privacy related reasons, personal archiving method is not practical. Therefore, the preferred method for this proposed solution is scraping.

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability of a given sample belonging to a particular class. Bayesian classifier is based on Bayes’ theorem. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computation involved and, in this sense, is considered “naive” [13]. Naïve Bayes method recommends hashtags by observing the content produced by the target user. In this paper it proposed to use Bayes model to estimate the probabilities of using different hashtags. Using this method, hashtags which are used by posts that has similar content can be identified [14].

The Natural Language Toolkit is a suite of program modules, data sets, tutorials and exercises, covering symbolic and statistical natural language processing. NL TK is written in Python and distributed under the GPL open source license. Over the past three years, NLTK has become popular in teaching and research [15]. In the proposed research application, Naïve Bayes classifier allows to classify the generated hashtags based on the analysis that given by the specific algorithms that trained using large training data-sets.

III. Research Gap

Currently, there are few hashtag generating applications which can fetch hashtags only for a given input text. The current platforms do not facilitate users to generate hashtags based on images, paragraphs and/or URLs. Hence the user has to manually search for hashtags through the existing applications. Existing applications has a collection of hashtags which rarely updates with time. As a result, a precise decision cannot be made whether the fetched hashtags are trending or not. Therefore, the hashtags offered by these applications are mostly outdated and have a less tendency for the user to achieve the expected reach for his publications. Furthermore the user is unable to sort out the suggested tags according to the targeted audience. This is a major disadvantage for corporate users that use social media as a mode of advertising platform for their purposes and digital marketers as well as the personals that are involved in social commerce. The next identified gap is the current applications are unable to guarantee the authenticity of the seller uploaded content. Seekmetrics, All-hashtags and Hashtagify are few names of above mentioned applications which can generate hashtags only for a small text input. When the text input is heavy, the accuracy of the performance drops to a recognizable rate.

Trending Hashtag Generator and Content Authenticator covers all the gaps identified in the existing applications mentioned above.

IV. Methodology

A. Analysis and Requirement Gathering

The analysis phase was focused on gathering information about the existing systems and analyzing the weaknesses and strengths of the respective systems which lead to the concept of developing the new system. Requirements of the new system were clearly understood during the analysis phase. The research team identified main users of the ‘Trending Hashtag Generator and Image Authenticator’ are social media marketers, promoters and their target audience. After conducting a survey, the research team found that more than 70% of online marketers use hashtags to promote their content and majority of them were not satisfied with their audience reach even after using social media platform audience tools. 60% of online marketers and promoters were using existing systems to find matching hashtags for their content. Authors came into a conclusion that an efficient system which analyzes the content and generate "real time trending hashtags" would assist the online marketers and promoters to get the audience reach they expect.

B. Implementation

The overall system was developed and built during this phase. The system architecture is mainly divided into four main components.

As depicted in Figure 1, authors are following a sequential approach to implement the proposed solution. Figure 1 also illustrates the type of output expected from the proposed solution.

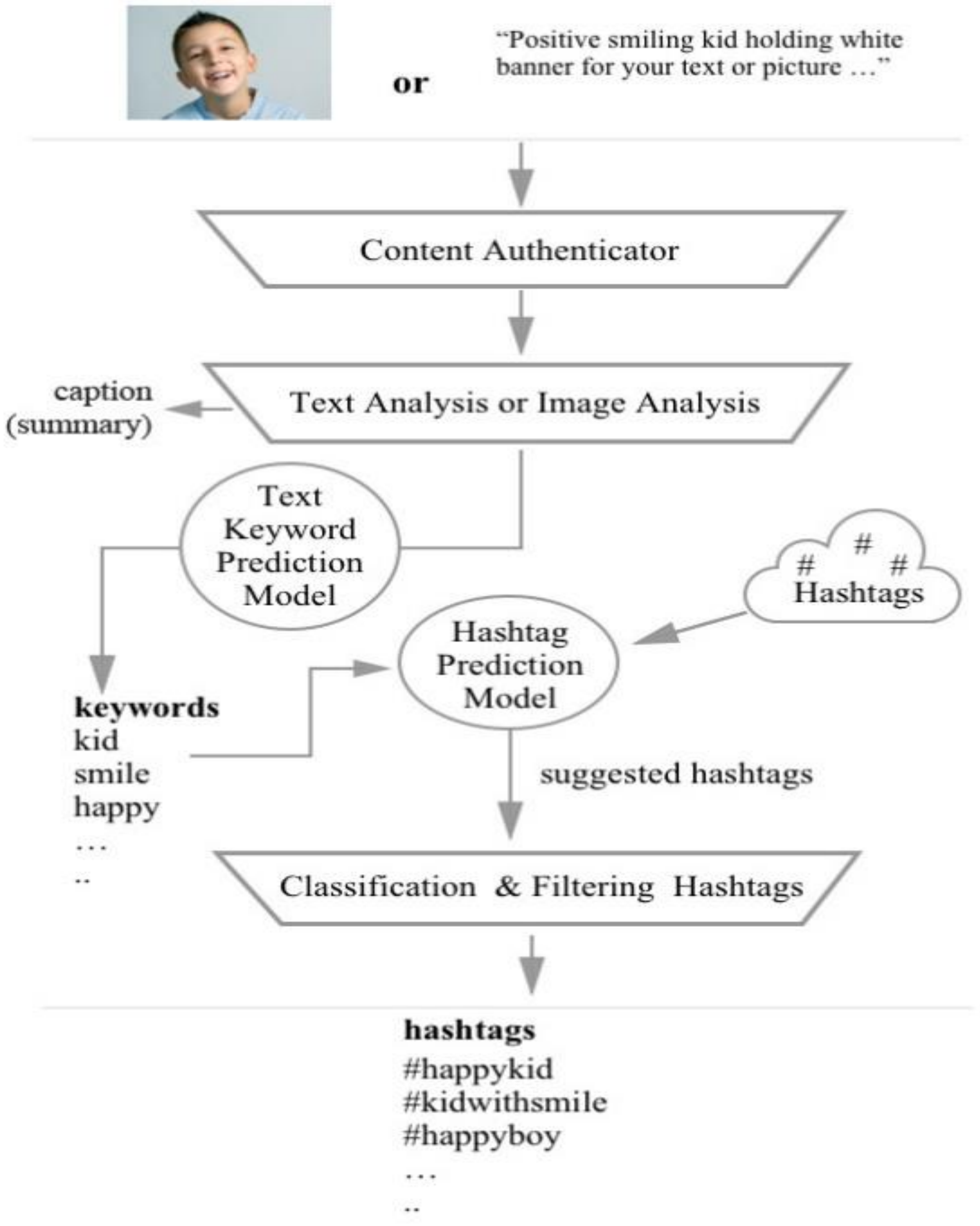

or
"Positive smiling kid holding white banner for your text or picture …"
Content Authenticator
Text Analysis or Image Analysis
caption (summary)
Text Keyword Prediction Model
# Hashtags
Hashtag Prediction Model
keywords
kid
smile
happy
…
..
suggested hashtags
Classification & Filtering Hashtags
hashtags
#happykid
#kidwithsmile
#happyboy
…
..

Fig. 1. System Architecture Diagram

C. System Components

1) Image Authentication using Metadata

In this process, the content authentication will be ensured with the usage of metadata. The metadata will be extracted with the aid of java libraries and subjected for an appraisal. The main metadata tags that will be extracted will be the make of the camera, model of the camera and the tags that will aid in locating the geographical location and the serial number. The appraisal will be done based on a defined criteria that will aid in finding the authenticity of the uploaded image. This criteria will be mainly based on the quality of the Metadata. Based on the appraisal, a rating will be provided to the user and it will be provided based on predefined assumptions. These assumptions are created based on the targeted audience which are social media influencers and commercial users. Once the user gets the rating, the user will be provided with an option to accept the rating or to request for a re-assessment and provide any justification if required if the rating is not in a satisfactory level. Once the rating is accepted the user can move on to the next phase.

2) Image Feature Extraction

The main objective of this process is to identify if the user uploaded images are already existing in the internet. The image feature extraction process is done based on the scale-invariant feature transformation which is used to detect and describe local features in images. Correspondingly speeded up robust feature techniques are used for object recognition, classification and image regression. Feature extraction is done using the images that are harvested from a web crawler and to the images that gets uploaded by the publisher. Once both features are extracted they will be compared to get a match. The feature extraction process will recognize interest points of the image and the RGB values will be compared. This matching process depends on the Euclidean distance between the interest points of the two images.

3) Text Analysis

This process is based on Natural Language Processing (NLP) and Machine Learning (ML). Python is used as the programming language. If the user inputs a URL which he needs to generate hashtags, the text content in the respective URL is fetched removing all the HTML markup elements and other unnecessary items. Using NLP libraries, stop words and punctuation marks are removed. The repetitive words are removed along with the canonical form of words. After above processes, a list of important keywords are generated. Nouns, verbs, pronouns etc. are separately identified. The nouns are then analyzed and automatically categorized into unique topics. This process is known as Topic Modeling or Latent Dirichlet Allocation (LDA). A unique ID is allocated to each topic and the relevant topic model is compared with a manually created dataset (1000 keywords and similar words related to each keyword). This process is carried out using cosine similarity algorithms which is commonly used in data science domain. The final result is a recommender system which is an application of machine learning which recommends related, most matching keywords for the generated words in LDA topic model. The generated keywords and recommended keywords are then passed in order to generate hashtags. The ultimate target in this phase is to automatically suggest many keywords as possible to user so that the system will generate various hashtags which are relevant to the user posted content. As a result, the user can get a broad and specific target audience's reach. Additionally, a meaningful summary is generated using term frequency–inverse document frequency (TF-IDF) which can be used as a caption when a user posts a post. If the user inputs a direct text, a similar process is carried out using NLP and ML techniques respectively. In this case, considering HTML elements is unnecessary.

4) Recommending Hashtags

The aim of this component is to recommend the best hashtags for the user content. A dataset which includes hashtags with popularity data is prepared for further processing. Social media public APIs and web scraping algorithms along with python is used to extract hashtags. After data extraction, data is cleaned and converted into the desired format. The respective dataset is divided into two datasets where the first dataset is to develop the solution and the other dataset is to test the model. With respect to image analysis process, since image classification is a broad research area that is not the scope of this research, Google Vision API is used to analyze images along with existing artificial intelligence and machine learning technologies. In this process, real time hashtags are taken from Instagram. In order to perform the operation, four factors are considered. Calculation of Cosine similarity of the hashtags and the keywords is one factor. To perform the task, a dataset has to be prepared manually to calculate the cosine similarity. The frequency of the hashtags are considered within a given period along with the like and comment count for the relevant hashtags. Then weights are assigned to indicate the relevancy of hashtags accordingly. Finally, hashtags are inserted into an equation to calculate the final score. For these processes, Gensim, Numpy, Tensorflow framework, Word2vec technologies are used. Since each iteration helps to perform the prediction more precisely, the training process is repeated.

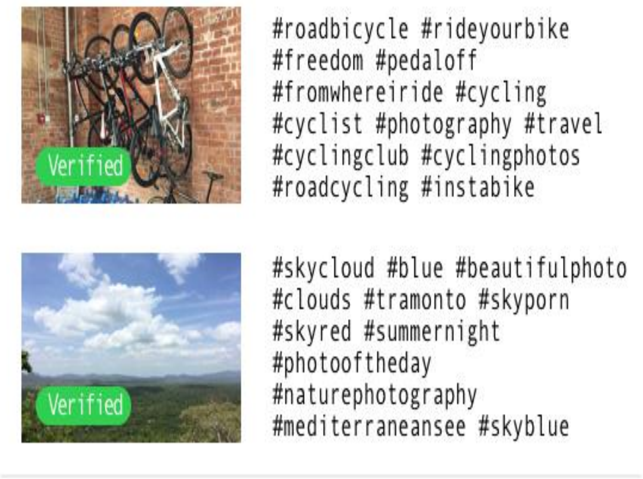
5) Classification and Filtering Hashtags

This process is based on Naïve Bayes Classifier and Machine Learning (ML). The generated hashtags are filtered into classifications to reach the ideal target audience. The selected filter option will analyze each and every hashtag in order to find the ideal sections of the hashtags. The related keywords of each hashtag is analyzed using a specific algorithm which uses Naïve Bayes Classifier to generate the results. Each category has its own unique algorithm in order to provide the classifying results, based on the specific target selection of the content uploader. The keyword analyzing algorithms are based on Naïve Bayes Classifier which will provide a probabilistic outcome of the related section of each hashtag. The classification process of the system is developed using ML which uses scikit-learn and natural language toolkit (NLTK). The classifying algorithms are trained using specific data models which increases the quality of the classifying result. According to the selected filter option, a unique classifier algorithm analyzes the keywords in order to classify the generated hashtags.

V. Result & Discussion

Hashtag Generator and Content Authenticator is a web based application that allows users to find the most popular hashtags for user specific content to get a recognizable amount of reach from a target audience with the content authentication service.

The images which go through the uploading process initially undergo the content authentication phase where the images will be authenticated with the aid of the metadata and feature extraction. These authentication processes are parallelly executing. Once the authentication is done, a tag will be shown to the audience indicating the authenticity of the images. Afterwards, the images will be forwarded to the image analysis component which generates a set of keywords relevant and unique to each uploaded image. If user enters a text content instead of an image, text will be analyzed and important keywords are analyzed along with a meaningful summary. Additionally, most related keywords are suggested to the user using the trained keywords model. In the next phase, keywords will be analyzed and relevant hashtags will be suggested using the trained hashtags model which consists of collected data of hashtags along with the popularity. Then the suggested hashtags will be classified and filtered in the last phase which allows the user to select his/her favorites based on the interest.



Verified

#roadbicycle #rideyourbike
#freedom #pedaloff
#fromwhereiride #cycling
#cyclist #photography #travel
#cyclingclub #cyclingphotos
#roadcycling #instabike

Verified

#skycloud #blue #beautifulphoto
#clouds #tramonto #skyporn
#skyred #summernight
#photooftheday
#naturephotography
#mediterraneansee #skyblue

Fig. 2. System Output

In the content authentication process, the selection of relevant images which needs to be harvested and exclude unwanted images can be considered as a limitation as it is a technical challenge to achieve 100% accuracy. Furthermore, a high processing speed is required in order to compare the features which are extracted between the images harvested and the image uploaded, which can also be considered as a technical limitation.

In the text analysis process, a data model (LDA topic model) is implemented to automatically separate keywords into specific groups. In this process, the accuracy of topic filtering is not 100% as the current LDA topic modelling has lot of room for improvement which is an active area of research.

A set of hashtags with popularity domain was initially required for the hashtag suggesting process but it was difficult to find an up-to-date hashtag dataset with the popularity details. The training process of hashtag model needed existing trending hashtag dataset models. The retrieval of some datasets which has the popularity domain from social media is increasingly becoming difficult due to the recent social media privacy policy changes. General Data Protection Regulation (GDPR) changes in the European Union is a clear example for such rule change.

In the hashtag categorizing process, though there are infinite methods of categorizations, only a limited number of specific categorizations are considered. Each and every selected category requires specific algorithms which need to be trained with unique data sets.

VI. Conclusion

The problem of accuracy and authenticity of the content published by online marketers is solved. A proper guidance can be provided to online sellers who don't have any idea about trending hashtags and reaching the proper target audience. The Metadata Extraction process successfully contributes towards the verification of the authenticity of the user uploaded content with the help of Image Feature Extraction process. Text Analysis phase generates a meaningful caption to user uploaded text content and will make sure the necessary keywords are generated and suggested at the same time which helps the seller to get more attention from the audience. The custom made dataset is built to suggest keywords which goes toe-to-toe with current marketing game on social media. Hashtag recommendation process identifies the image and/or the keywords supplied from the Text Analysis phase and generates relevant, real time trending hashtags. The generated hashtags are then categorized based on different attributes and the categorized hashtags are suggested to the user (online marketers).

VII. Future Works

In the text analysis process, the recommender system uses a manually built dataset with 1000 keywords which are limited to the following domains; fashion, nature and travel. The dataset can be extended to several other domains as future works in order to cater user inputs from other domains.

General recommendations to those who are willing to develop this system further are as follows:

- Expand the API enabling the access to 3rd party users.
- Develop the accuracy of the topic model (LDA topic model) which suggests related keywords.
- Develop the advanced popularity prediction mechanism for the suggested hashtags.
- Expand the system in to mobile platforms as Android and IOS.
- Embed the personalization in to the system. Develop models to suggest hashtags depending on the person's past hashtag usage patterns.
- Expand the image authentication for wide range of images on the internet.

References

[1] 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/. [Accessed: 06- May-2018].

[2] W. Hussein, M. Salama and O. Ibrahim, "Image Processing Based Signature Verification Technique to Reduce Fraud in Financial Institutions", Matec conferences.org, 2016. [Online]. Available: https://www.matec-conferences.org/articles/matecconf/pdf/2016/39/matecconf_cscc2016_05004.pdf. [Accessed: 10- May- 2018].

[3] B. Yohanes and H. Wardana, "Focused Crawler Optimization Using Genetic Algorithm", Telkomnika.ee.uad.ac.id, 2018. [Online]. Available: http://telkomnika.ee.uad.ac.id/n9/files/Vol.9No.3Des11/1SS-AI9.3.12.11.01.pdf. [Accessed: 12- May- 2018].

[4] Chowdhury, G. (2005). Natural language processing. [online] Available at: https://onlinelibrary.wiley.com/doi/full/10.1002/aris.1440370103 [Accessed 15 Mar. 2018].

[5] Li, S. (2018). Topic Modeling and Latent Dirichlet Allocation (LDA) in Python. [online] Towards Data Science. Available at: https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24 [Accessed 3 Aug. 2018].

[6] Mo, Y., Kontonatsios, G. and Ananiadou, S. (2015). Supporting systematic reviews using LDA-based document representations. [online] Available at:

https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-015-0117-0 [Accessed 4 Aug. 2018]

[7] Liu, L., Tang, L., Dong, W., Yao, S. and Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. [online] An overview of topic modeling and its current applications in bioinformatics. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5028368/ [Accessed 5 Aug. 2018].

[8] Z. Ding, Q. Zhang, and X. Huang, "Automatic Hashtag Recommendation for Microblogs using Topic-Specific Translation Model.," in COLING (Posters) , 2012, pp. 265–274

[9] Z. Ding, X. Qiu, Q. Zhang, and X. Huang, "Learning Topical Translation Model for Microblog Hashtag Suggestion.," in IJCAI , 2013, pp. 2078–2084.

[10] E. Zangerle, W. Gassler, and G. Specht, "Recommending\#-Tags in Twitter," in Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings , 2011, vol. 730, pp. 67–78.

[11] J. Weston, S. Chopra, and K. Adams, "#TagSpace: Semantic Embeddings from Hashtags.," in EMNLP , 2014, pp. 1822–1827.

[12] E. Denton, J. Weston, M. Paluri, L. D. Bourdev, and R. Fergus, "User Conditional Hashtag Prediction for Images.," in KDD , 2015, pp. 1731–1740.

[13] Ming K. 2007 Naive Bayesian Classifier [online] Available at: http://cis.poly.edu/~mleung/FRE7851/f07%%202FnaiveBayesianClassifier.pdf

[Accessed 04 Aug. 2018]

[14] Su Mon 2011 On Recommending Hashtags in Twitter Networks [online] Available at: https://link.springer.com/chapter/10.1007/978-3-642-35386-4_25 [Accessed 16 Mar. 2018]

[15] Steven Bird 2002 NLTK: The Natural Language Toolkit [online] Available at: https://dl.acm.org/citation.cfm?id=1118117 [Accessed 05 Aug. 2018]