Implementation Paper On Social Media Data Analysis

Prof. Ram Joshi, Prof Mehzabin Shaikh, Aditi Patil, Dipti Patil, Rutuja Sonawane, Shivani Suryawanshi

Aditi Patil: Student, IT department, JSPM's Rajarshi Shahu College of Engineering, Tathwade, Pune-Savitribai Phule University-India

Dipti Patil: Student, IT department, JSPM's Rajarshi Shahu College of Engineering, Tathwade, Pune-Savitribai Phule University-India

Shivani Suryawanshi: Student, IT department, JSPM's Rajarshi Shahu College of Engineering, Tathwade, Pune- Savitribai Phule University-India

Rutuja Sonawane: Student, IT department, JSPM's Rajarshi Shahu College of Engineering, Tathwade, Pune-Savitribai Phule University-India

Mehzabin Shaikh: Professor, Dept. of Information Technology, JSPM's

Abstract: As the ever-growing segment of people uses social media in their daily lives, social media is analyzed in many different fields. The process of analyzing social media involves four distinct stages, data acquisition, collection, optimization, and analysis. Social media analytics is the process of collecting hidden information from social data - both formal and informal - to allow informed decision-making.

Recommendation systems are widely used on the web to recommend products and services to users. Most e-commerce sites have such systems. But here social media platform is something we plan to do data analysis on. We know somehow Instagram, Facebook, etc are being used for this, but photos or via posts as a dataset. Here description using hashtags is something new we are going to try as a source for recommending products. As the hashtag labels used on social media sites make it easy to find specific themes or content and are used not only to describe the visual content of an image but also serve other functions falling under the meta-communicative use or describing things.

This paper proposes a one-of-a-kind recommendation system through data analysis on social media where we are extracting new datasets like keywords including hashtags, which are further used to recommend similar products.

Keywords: Hashtags, pre-processing, web scraping, classification, clustering, feature Extraction K-Nearest Neighbour (KNN).

1. INTRODUCTION

A product recommendation engine is an answer that permits advertisers to offer their clients important item proposals in real-time [8]. As powerful data filtering tools, recommendation systems use algorithms and data analysis techniques to endorse the most pertinent product/items to a specific user. The fundamental point of any recommendation engine is to animate requests and effectively connect with clientsa component of an eCommerce personalization strategy, recommendation engines dynamically populate different items onto websites, apps, or emails, thus enhancing the client's experience. These sorts of shifted and omnichannel suggestions are made based on multiple data points such as customer preferences, past transaction history, attributes, or situational context. Recommender systems can be used across multiple verticals, for example, online business, entertainment, portable applications, education, and more. To summarize, a recommendation engine can be useful in any circumstance where there is a need to offer clients customized ideas and guidance [9]. Whereas this paper works in quite a unique manner as without any user-item purchase history, a search engine-based recommendation system can be designed for users. The product recommendations can be based on textual clustering analysis given in the description which is retrieved from social media handles.

Web scraping- When it comes to the dataset from social media like Instagram, it's hard to arrange for the dataset we need [11]. As Instagram is a dynamic site. After going through various methods, web scraping is something we found, which is a collection of R scripts that can be used to crawl public Instagram data without the need to have access to the official API. Its functionality is as same as compared to what is possible using the official API. However, it seems to be the only option for non-developers to gather and analyze Instagram data.

2. LITERATURE SURVEY

[1]. "A Social Network-Based Recommender System (SNRS)"-2010

The paper presented a social network-based recommender system (SNRS) which makes recommendations by considering a user's own preference, an item's general acceptance, and influence from friends. In particular, it proposed to model the correlations between immediate friends with the histogram of friend's rating differences. The influences and impressions from distant friends are also considered in an iterative

[2]." Recommender Systems in E-Commerce"-2014

Recommender systems allow e-commerce sites to be highly customizable for the user and buyer. The paper includes recommendations using association rules are generated based on previous transactions the user has already displayed interest in. Collaborative filtering allows the active user to get a recommendation based on products that users with similar interests have purchased and rated positively, and by using the active user's previous ratings and transaction history to build a model that provides a new set of similar products. Content-based filtering analyses the user's profile and preferences with the database to find products that are of interest and line up with the active user and present them.

[3]. "Instagram Post Data Analysis"-2016

Due to the spread of the Internet, social platforms have become large information pools. The paper has completed a visualization project on Instagram data. This study the relationship between the likes and the hashtags, location, and filter, it shows the relationship of Instagram filter data with location and number of likes to give users filter suggestions on achieving more likes based on their location. It likewise analyses the popular hashtags in different locations to show visual culture contrasts between different cities. It proposes a recommendation system, which can give general suggestions on the choice of filters based on location

[4]. "Language in Our Time: An Empirical Analysis of Hashtags"-2019

Hashtags in online social networks have acquired tremendous popularity during the past five years. The resulting large quantity of data has provided a new lens into modern society.t. The paper demonstrates hashtags that are more consistently shared among users, as measured by the proposed hashtag entropy, are less inclined to semantic displacement. In the end, it proposes a bipartite graph implanting model to summarize users' hashtag profiles, and depending on these profiles to perform friendship prediction Evaluation results show that this methodology achieves an effective prediction with AUC (area under the ROC curve) above 0.8 which demonstrates the strong social signals possessed in hashtags.

[5]. "Sentiment Analysis of Twitter Data"-2019

Nowadays, people from all around the globe use social media sites to share information. Twitter for instance is a platform in which users send, read posts known as 'tweets' and interact with different communities. The paper proposed model used several algorithms to enhance the accuracy of classifying tweets as positive, negative, and neutral. In this paper, data extracted directly from Twitter API were used to train and test the models. A lexicon-based classifier used a manually created lexicon to find the sentiment of each tweet.

3. METHODOLOGY

The System Architecture Diagram for recommendation using Data analysis of social media is shown in figure 1

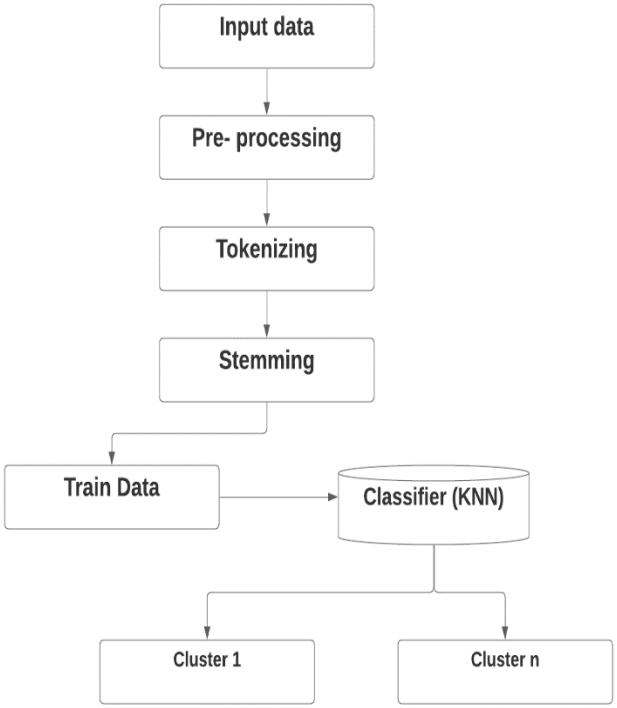


Fig 1: SYSTEM ARCHITECTURE

3.1 Pre-processing

The data input contains files including activity information of all the users from different social media handles like Instagram, Facebook, etc. Further, we have to extract the description column from data input which will be in the form of text, hashtags, description, etc.

	ID	Img_URL	Likes	Owner	Description	Date
1	CPDuFa5Av9M	https://instagram.fpnq2-1.f	0	14722909543	Best Selling Adjustable Stereo Gamer Earphone Headphones At 5	########
2	CPDtkkAl2Il	https://instagram.fpnq2-1.f	6	349020262	<U+041F><U+0435><U+0440><U+0432><U+044B><U+0435> TWS	########
3	CPDtHdsL_Ca	https://instagram.fpnq2-1.f	8	34496907150	Ace your grind with the all new sweat proof Rapz H1. Always stay	########
4	CPDspmolPhw	https://instagram.fpnq2-1.f	1	40049127753	Check out the best wireless earbuds available in India 2021. LINK	########
5	CPDrElqF2Ow	https://instagram.fpnq2-1.f	8	42801466013	REAL ME AIR PRO + <U+0001F60D> AVAILABLE IN THREE COLOUR	########
6	CPDp7FJNnd-	https://instagram.fpnq2-1.f	2	2293595208	Let's step out of the office, put on your earbuds to play your favo	########
7	CPDp5xFDMQZ	https://instagram.fpnq2-1.f	1	37529958004	Donâ€™t just listen to music, feel it. JAMKIX Pro provides you the	########
8	CPDp6n3ljfr	https://instagram.fpnq2-1.f	1	45957164875	It is a good headset at a reasonable price, allowing you to listen t	########
9	CPDp3YYgCUJ	https://instagram.fpnq2-1.f	0	47239617358	<U+0001F69B> SÃ³ hoje FRETE GRÃTIS quase todo Brasil! Headpl	########
10	CPDpCpPFbL7	https://instagram.fpnq2-1.f	5	47787194120	1+tws <U+0001F4AB> Price - 9000 rupees only<U+2764><U+FE0F>	########
11	CPDo4lsBl0x	https://instagram.fpnq2-1.f	2	47326148026	Fone de ouvido digital bluetooth R$ 120,00 Fone Ouvido Digital T	########
12	CPDofITBI5Y	https://instagram.fpnq2-1.f	3	47326148026	Fone de ouvido bluetooth JBL. R$ 120,00. Fone de Ouvido Headp	########
13	CPDoHhlBcF4	https://instagram.fpnq2-1.f	0	47725696314	#Realme #Airneobuds #bluetoothheadphones	########
14	CPDn6SSBQ5D	https://instagram.fpnq2-1.f	2	47383942697	Now available Premium quality airpods pro that are compatible v	########
15	CPDn497LAhf	https://instagram.fpnq2-1.f	13	11919065651	Amani Wireless Headphones ASP BT-5760: Wireless Bluetooth Or	########
16	CPDnAZEBdn9	https://instagram.fpnq2-1.f	5	47853189141	REALME AWESOME COMBO <U+0001F525> *REALME SMART W	########
17	CPDmd0grqjE	https://instagram.fpnq2-1.f	1	2233778438	Have both of your hands to your own disposal with a wireless de	########
18	CPDkvT4AxQu	https://instagram.fpnq2-1.f	40	9312129788	Extended work hours at home calls for chill thereafter. Shop #bl	########
19	CPDkUzxna1-	https://instagram.fpnq2-1.f	7	7383022758	Customize Photo with Bluetooth speaker Description An Awesor	########
20	CPDjsDCoUNv	https://instagram.fpnq2-1.f	0	6618811559	Posh Headphone Starting From - 399 Follow us @posh_ace #hea	########

Fig 2: Data set of Instagram

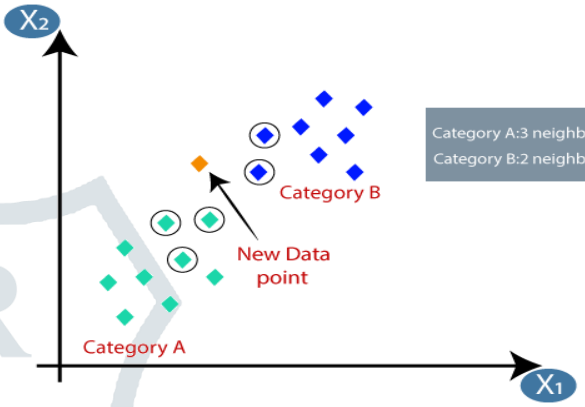
E.g., Fig 2 represents sample data extracted from Instagram which includes a number of likes, captions in the form of description, date of the post, etc. To convert these sentences of description to words we use Tokenization for removing unnecessary punctuation, tags [10]. That further is followed by removing stop words i.e., Frequent words such as “the”, “is”, etc. that do not have specific semantic. And lastly Stemming is performed where words are reduced to a root by removing inflection through dropping unnecessary characters, usually a suffix or prefix.

3.2 Classification.

The training data is trained by using machine learning algorithms i.e., K- Nearest Neighbour (KNN):

K-Nearest Neighbors algorithm (or KNN) is quite possibly the most utilized learning algorithm because of its effortlessness. KNN is a non-parametric algorithm. It utilizes information with several classes to anticipate the grouping of the new sample point. While carrying out Discriminant examination when some dependable parametric controls of probability densities are not known or found challenging to understand this classification method was developed to perform such calculations. The exact location of the K-nearest neighbors should be decided with the help of the training dataset. To discover how close every individual of the training dataset is from the target is to be analyzed, we utilize Euclidean distance. Discovery of the k-nearest neighbors and allocating the group to the row that is being inspected. Now repeat the technique for the rows outstanding in the target set. We can also select the maximum value of K in this software after that the software automatically builds a parallel model on the values of k up to the maximum that specifies the value. The best score achieved of k between 1 and the given value is chosen that helps to build parallel models on all values of k up to the extreme identified, an incentive for which k=10 was selected and scoring is done using the finest models from the available ones. At last, the data required for classification is entered

The K-Nearest Neighbor applies the Euclidean distance formula as shown in Fig 3. The Euclidean distance between training and a test tuple can be written as follows: Let Xi be an input tuple with p features (xi1, xi2, …, xip). Let n be the



total number of input tuples (i = 1, 2,3,4 …, n), while p be the total number of features (j = 1, 2, …, p). The Euclidean distance between Tuple Xi and Xt (t = 1, 2, …, n) can be derived as $d(x_i, x_t) = \sqrt{(x_{i_1} - x_{t^1})^2 + (x_{i^2} - x_{t^2})^2 + \cdots + (x_{i_P} - x_{t_P})^2}$)

i.e.,$\text{dis}(x, x_2) = \sqrt{\Sigma_{i=1}^{n}(x_1 i^{-x_2} i)^2}$

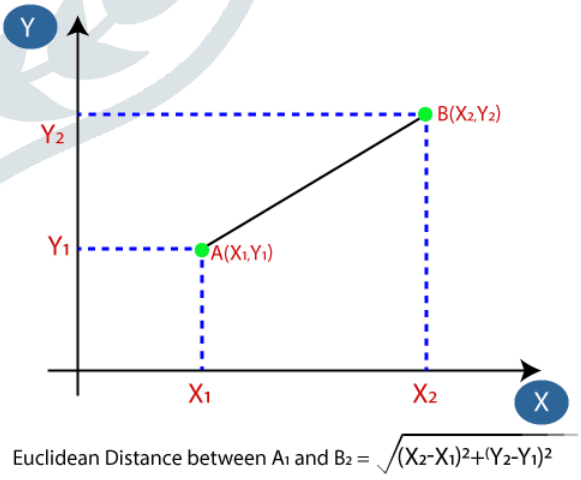


Fig 3

In K-NN, classification all adjoining points that are closest to the test tuple are encapsulated and recommendation is made dependent on the closest distance to the test tuple as shown in fig 4 and 5, this can be defined as follows:

Let *C* be the predicted class

$$C_i = \{x \in C_p; d(x, x_i) \leq d(x, x_m), i \# m\}$$

For e.g., Fig. 6 shows the formation of clusters based on the input file mention in Fig 2

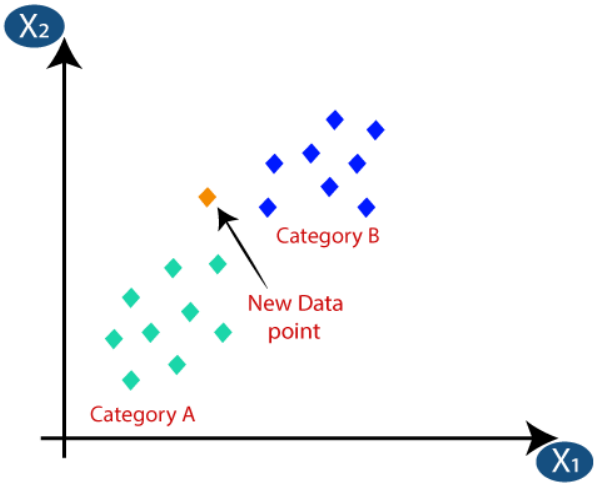


Fig 4

.

Fig 5

3.3 Clustering

Clustering refers to the process of automatically grouping data points with comparative attributes and assigning them to "clusters." Some use cases for clustering include Recommender systems (grouping users with similar viewing patterns on Netflix, to suggest comparative content). The main focus of this algorithm is discovering the groups in the data with that number of groups that represent the variable K. This algorithm iteratively allocating the k groups to the point. Data points here are clustered based on a feature of similarity.

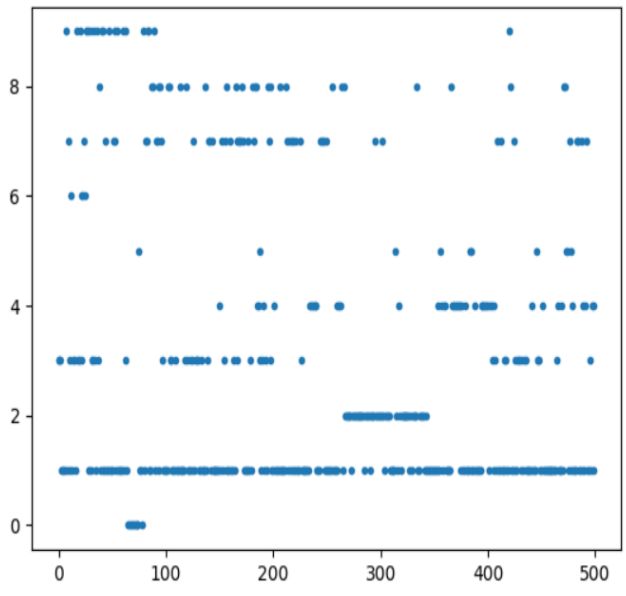


Fig 6

3.4. Predicting

To enhance the accuracy of recommendation results, we use the iterative clustering method to cluster users before recommending [6]. As we know that users in a cluster will have similar interests, thus, if a user is using say hashtags related to some specific content in feed posts, IGTV or stories, our system will generate suggestion based on those hashtags and description used in the caption and suggest a product which will be related to the content the user might be interested in. And hence we recommend the best top 10 suggestions to the user.

Fig 7

Here e.g. above fig 7 represents the output in the form of formation of an array of clusters and depending on that predication is made but here, we not only recommend products but our algorithm works in such a way that it generates URLs of recommended products via amazon website and these URLs will be the output that we will be predicting for the user.

4. DATABASE AND SOFTWARE USED:

4.1 Database used

When it comes to dataset from social media like Instagram, it is hard to arrange for dataset we need. As Instagram is a dynamic site. After going through various methods, web scrapping is something we found, which is a collection of Rscripts that can be used to crawl public Instagram data without the need to have access to the official API. It's functionality is as same as compared to what is possible using the official API.

```
#----------------------------
#Get New Posts from Instagram
#----------------------------
getNewPosts <- function(index){
  print("getNewPosts function called")
  url_next <- str_glue("{url_start}&max_id={end_cursor}")
  json <- fromJSON(url_next)
  edge_hashtag_to_media <- json$graphql$hashtag$edge_hashtag_to_media
  end_cursor <- edge_hashtag_to_media$page_info$end_cursor
  posts <- edge_hashtag_to_media$edges$node
  assign("end_cursor", end_cursor, envir = .GlobalEnv)
  assign("posts", posts, envir = .GlobalEnv)
  print(index)
  Sys.sleep(1)
  extractInfo(index)
}

#Start the Madness
extractInfo(index)

#----------------------------
#Export Dataframe to CSV()
#----------------------------

table <- do.call(rbind.data.frame, Map('c', post_id, post_img_url, post_likes, post_owner, post_text, post_time))
colnames(table) <- c("ID", "Img_URL", "Likes", "Owner", "Description", "Date")
time <- Sys.time()
filename <- str_glue("{hashtag}.csv")
write.csv(table, filename, fileEncoding = "UTF-8")


#May run first to set TZ
Sys.setenv(TZ="Europe/Berlin")
Sys.getenv("TZ")
```

Fig 8 : Code for Data Extraction

4.2 Software used

To implement this we have used Eclipse, JDK, Apache-Tomcat software. For frontend we have used HTML, CSS, Javascript and for algorithm we have used python because KNN algorithm is only available in python language.

5. CONCLUSION:

The proposed system is highly efficient in terms of complexity. Provides highly accurate results or recommendations to system users. The result of our experiment is a real-time referral machine operated by the KNN classification model implemented with the Euclidean remote system that is effective and provides a good and accurate clustering and suggestions based on the client's immediate need for information without any user-item purchase history by using social media data analysis. To summarize, our system performs product recommendations based on textual clustering analysis given in the description below-posted picture (i.e., the caption area) which is retrieved from social media handles say Instagram or Facebook, and it not only predicts products but, as an output, it gives a downloadable output file which consists of URLs from the website like Amazon of predicted products. This means we can recommend the user, products based on their caption information unlike recommending based on the user's history of information.

6. REFERENCES:

1. V. Krishnaiah, G. Narasimha, N. Subhash Chandra, "Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review" IJCA 2016.
2. "A Social Network-Based Recommender System (SNRS)"-2010
3. "Recommender Systems in E-Commerce"-2014
4. Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method-2016 https://www.sciencedirect.com/science/article/pii/S221083271400026X
5. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
6. Lisa Posch, Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2013. Meaning as Collective Use: Predicting Semantic Hashtag Categories on Twitter. In Proceedings of the 2013 International Conference on World Wide Web (WWW). ACM, 621–628.
7. Yang Zhang, Mathias Humbert, Tahleen Rahman, Cheng-Te Li, Jun Pang, and Michael Backes. 2018. Tagvisor: A Privacy Advisor for Sharing Hashtags. In Proceedings of the 2018 Web Conference (WWW). ACM, 287–296.
8. Luca Maria Aiello and Nicola Barbieri. 2017. Evolution of Ego-networks in Social Media with Link Recommendations. In Proceedings of the 2017 ACM International Conference on Web Search and Data Mining (WSDM). ACM, 111–120
9. Jisun An and Ingmar Weber. 2016. #greysanatomy versus #yankees: Demographics and Hashtag Use on Twitter. In Proceedings of the 2016 International Conference on Weblogs and Social Media (ICWSM). AAAI, 523–526.
10. [10] F. Kunneman, C. Liebrecht, A. van den Bosch the (un)predictability of emotional hashtags in Twitter Proceedings of the 5th Workshop on Language Analysis for social media, (LASM), Association for Computational Linguistics (2014), pp. 26-34CrossRefView Record in Scopus Google Scholar
11. https://github.com/JonasSchroeder/InstaCrawlR