Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 4, 2020

Review Article

MEME CHAT: AN INNOVATIVE SOCIAL MEDIA PLATFORM FOR CONTENT MONETIZATION & CORPORATE OUTREACH USING OCR & NLP

Dr.K.M. Umamaheswari[1], Taaran Chanana[2], Kyle Fernandes[3]

[1]Assistant Professor, Department of Computer Science & Engineering, SRM Insitute of Science & Technology, Kattankulathur, Tamil Nadu, India. umamahek@srmist.edu.in
[2]IV Year B.Tech (CSE), SRM Insitute of Science & Technology, Kattankulathur, Tamil Nadu, India. ctaaran@gmail.com
[3]IV Year B.Tech (CSE), SRM Insitute of Science & Technology, Kattankulathur, Tamil Nadu, India. kyleferna@gmail.com

Received: 07.12.2019 *Revised: 13.01.2020* *Accepted: 25.02.2020*

Abstract
Social media has boomed in recent years, along with platforms for user generated content. Major social network platforms like Google and Facebook use a singular advertisement mediation service, where ads are provided by corporates to target users/customers. Facebook analyses and uses a user's meta data and content consuming behavior to present the best suitable commercial. We, in this work, show an innovative approach where advertisement mediation is done using user's own generated content, in the form of memes on a corporate. Memes are chosen as a medium due to their supreme ability to penetrate the internet and get viral quickly. Because the content is generated by the user, who is also the target of the advertisement, the content is well understood, received and appreciated in comparison of content used in other approaches like Facebook audience network, Google Adsense or Twitter MoPub. Using this method, users are renumerated monetarily for the content they post, which also becomes the advertisements that will be later used by corporates. We track user generated content, for subscribed corporates using Tesseract algorithm for optical character recognition along with Vader NLP. for NLP. and provide results that strongly suggest why the method suggested in this article is useful and advantageous compared to other methods.

keywords*:* Online Advertizements, Social Network based Marketing, Optical Character Recognition.

INTRODUCTION

In current day and age, with internet on the rise and over 600 million users using it on a daily basis from India alone; this has opened up many opportunities for brands and companies to market their products which range from products, services to movies. The platform provides effective marketing solution to these brands and companies through the use of OCR and NLP to enable progress tracking.

We provide a platform to both Brands and Users to mutually benefit out of it with the help of OCR and NLP technologies. Firstly, the 250,000 app users are allowed to make humorous images on those topics - the Brand's engagement is then tracked using OCR. (Optical Character Recognition) and Sentiment Analysis. This data is analyzed via our algorithm and if it's found to be in relation with the Brand, the user gets monetary renumeration. A trained and tested model which is a use case of Tesseract and modern NLP with a reference dictionary is used to understand the text and contents inside the images, text comments and conversations between users which is tracked at every point. It is shown that the technology used in this article solves the issue of efficient content tracking pertaining to brand more efficiently than other advertising platforms. We provide detailed insights and reports of each action placed on their product. This improves performance tracking and indirectly, client relation by a significant margin.

The rest of the paper is split up as : Section II reviews the common Advertising strategies used by most social network-ing platforms which comprises of majorly to any Social net-working companies revenue, it also reviews the best practices and approach to implement our strategy using Tesseract OCR and Sentiment Analysis; Section III describes our proposed methodology as well as our innovation and improvements to implement the same; Section IV deals with the algorithm of of how the entire methodology functions; Section V with analysis of obtained results and Section VI provides a brief conclusion to the paper.

LITERATURE SURVEY

There is a role of social media platforms in creating a new type of more immersive user experience that provides a ’sense of belongingness’ to the user and how corporations can use this ’sense of belongingness’ to better target the user. A discussion about the significance of the role the ’virtual community’ created by social media plays is as explained in [1].

Importance of user motivation, types of user motivation, which are ’psychological well being’ and ’online social capi-tal’ are shown in [2]. It also discusses types of social media marketing, ’interactive digital advertising’ and ’virtual brand community’. The work consists of a survey conducted on 502 Facebook users from Taiwan. Results of the survey reveal that users respond to both types of social media marketing differently. It also states how in both of these cases, user motivation for using social network had influenced the recorded user response in the survey.

Perceived negative effects of intrusiveness in digital media based advertisement is documented in [3]. In this study, perceived negative effects of intrusiveness are connected to challenges in marketing strategy building. This study also provides extant thinking in this area. This work becomes a bases for further study in the domain of strategic marketing.

[4] provides detailed perceptions of the ones who practice digital media based marketing and advertizing. This includes in depth interviews of 21 key stakeholders in this market. Analysing these interviews, the authors found the key trends and prepositions that are relevant to advertising in this fashion which relate to

better formulation and management of digital advertising and marketing.

[5] Processing follows a traditional step-by-step system, but some of the stages in their day were rare, and probably even to this day remain so. The first step is an overview of linked components in which descriptions of the components are processed. This was a computationally expensive design choice at the time, but had a significant advantage: it is easy to detect inverse text and recognize it as easily as black-on-white text by analyzing the nesting outlines, and the number of outlines for mother and grandmother. Certainly Tesseract was the first OCR engine that could handle white-on-black text so trivially. At this stage outlines are gathered together into Blobs, simply by nesting. Blobs are grouped into lines of text, and lines and regions for fixed pitch or proportional text are evaluated. Text lines are fragmented into words differently depending on the nature of spacing between characters. Specific pitch text is instantly chopped by cells with characters. Proportional text is broken into terms using certain spaces and spaces that are flippant. Recognition then progresses as a two-pass process An attempt is made in the first pass to identify each word in turn. Every word that is appropriate shall be passed as training data to an adaptive classifier. The adaptive classifier then gets an opportunity to interpret text lower down the page more accurately.

[6] Tesseract started out as a PhD research project [2] in HP Labs, Bristol, and gained popularity as a probable software for HP's line of flatbed scanners. The reason behind the initiative was that OCR engines at that time were still in their infancy, and failed miserably on anything however the best first-rate print. After a joint initiative among HP Labs Bristol, and HP's scanner division in Colorado, Tesseract had a widespread lead in accuracy over the other OCR engines, however it did not turn into a product. The next level of its development started back in HP Labs Bristol to use OCR for compression.

[7] VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. This helps in deciding the positive or negative factor within text.

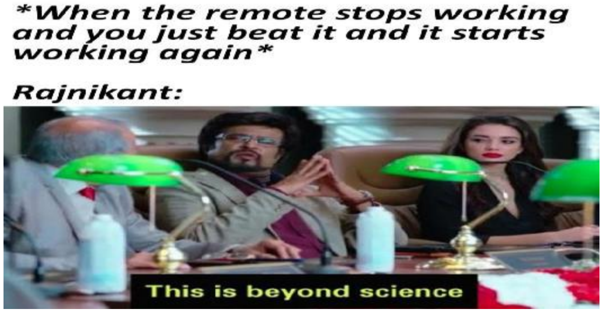

When the remote stops working and you just beat it and it starts working again
Rajnikant:
This is beyond science

Fig. 1: Example Meme

PROPOSED METHODOLOGY

The crux of the methodology focuses on extracting text from Images (memes), captions and comments, which fall under the category of UGC(user generated content). This approach uses an optimized version of Tessarect to extract text from images.

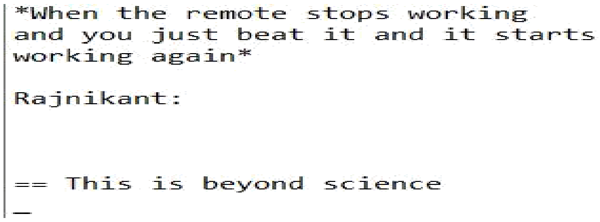
Then, a POS tagger (POS tagger), as the one suggested in [8], which uses the Penn-Treebank tagset, is used to process the unstructured text to get relevant keywords. Since India is predominantly Hindi speaking, The approach takes Hindi into consideration using Google Translate on the text obtained if it is in Hindi. The database stores these indexed keywords, along with keywords suggested by the brand.

We then further use Sentiment Analysis, as suggested in [7] is used on the comments, caption, and in Image text to understand the flow and direction of the brand conversation happening between users. Content that violates brand guide-lines are hidden, and content that promotes the brand are showcased. Users that make content that is Pro the brands guidelines and garners more positive comments, gets paid more. This is a never before seen marketing model where each individual user is a discreet agent and plays a major role in spreading awareness for a brand through word of mouth.

Tesseract OCR

Tesseract is an open-source optical character recognition engine, one of the most successful and best performing OCR libraries. OCR is using artificial intelligence to look for text and to detect it on images. Tesseract uses pixels, shapes, words and sentences to identify representations. This uses twostep approach, which calls adaptive recognition. For character recognition, it includes one data stage, then the second stage to satisfy some letters, it was not insured in, letters that can suit the meaning of the word or sentence.

An example of the same is given as follows. Consider the meme given in figure (1).



```
*When the remote stops working
and you just beat it and it starts
working again*

Rajnikant:



== This is beyond science

_
```

Fig. 2: Text as extracted by Tesseract OCR

For the above given meme, the tesseract OCR. algorithm derives the text as follows:

1. De-warping: Digital Image De-warping is the technique which deals with the geometric image transformations. The main challenges of camera image research is tackling the inconsistencies of cover and perspective. Given the prevalence of dewarping methods, the performance evaluation method[9] does not have a specific algorithm, with most of the evaluation being performed to focus on pleasing visual experiences. Our foremost goal is to recognize user generated textual content placed on snap shots in addition to pictures with text which can be essentially bounded volumes captured by using a virtual digicam and suffer from nonlinear warp. The proposed technique is carried out on grey scale document image and is based on numerous distinct steps the usage of an adaptive document image binarization[14], finally, a entire recuperation of the unique grayscale warped photograph guided by way of the binary dewarping which accomplishes the use of co-ordinate transform model in which the goal is to generate a transformation to flatten the warped photograph to its authentic shape. The transformation is a mapping from the curved coordinate system to a Cartesian coordinate system. Once a curved coordinate net is set up on the distorted photograph. The transformation may be done in steps: First, the curved net is stretched to a straight one, and then adjusted to a well-proportioned square net.
2. Binarization: Binarization is the first step for most document image analysis systems and refers to convert the image to a binary image on a grayscale. Image binarization transforms the image to a black and white image of up to 256 gray levels. Survey showed that global thresholding due to illumination variations is not appropriate for camera-captured images, so we suggested locally adaptive

thresholding methods that are robust to illumination variations. The simplest way to use binarization of images is to pick a threshold value, and mark all pixels with values above the threshold as white and all other pixels below that value as black.

3. Other Steps: Other steps taken include fixing dpi (dots per inch), text size, text lines and illumination of the image.

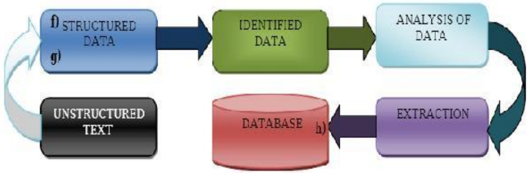


Fig. 3: Flow of processes in Text Mining

Text Mining

As mentioned in the previous sections, the proposed ap-proach uses sentiment analysis, a subfield in NLP (Natural Language Proccessing) to track and detect the user responses and through it, the effect of the generated content to provide and assess whether the user should be awarded monetary renu-meration. This work uses VADER (Valence Aware Dictionary and sEntiment Reasoner) for the same. Text mining is an important pre-processing step in the same.

The main fundamental steps involved in text mining are:

1) Gathering unstructured data from our dataset of meme images.
2) Detect and remove anomalies from data by conducting pre-processing and cleansing operations. Data cleansing lets you get and retain the valuable information hidden within the data and to help identify the classification of specific words.
3) Convert all the relevant information extracted from un-structured data into structured formats.
4) Analyze and record data.
5) Store all the valuable information into a secure database.

BRAND OUTREACH AND NLP CONNECTION

With the help of Vader Sentiment and OCR, once a person creates a meme and posts it, the following scenario takes place. For eg., let us consider the meme used in figure (1) and the text extracted from it as figure (2). From the same, using the POS. tagger from [8] which uses a Penn Treebank tagset, the keywords are extracted from the captions and the in image text as follows: remote,rajnikant,science.

These keywords are stored in a table which will have a simple schema pid(ID of that particular post), keyword. We then also have Keywords that would be given to us by the brand for which the ad. is being created. Let these keywords be as follows: Rajnikant,Akshay Kumar, Movie, Bollywood etc. (Keywords relevant to the topic being promoted).

If a meme contains those keywords, we link that content to the brand and brand gets notified. The user who created that meme gets renumerated just for creating that meme and the meme is then displayed on the main feed of the platform.

Now, assume we receive comments on that same meme such as follows:

A comments ”wow! Rajnikanth Played a great role”
B comments ”it was a boring movie!”
C comments ”bad movie! this meme is a lie.”

Based on the comments on the meme that was posted, if a users get more comments which bring out the reaction ’A’ the user will be re-numerated more, as compared to if the content brought out reactions such as ’B’ or ’C’. The positivity or impact or the sentiment of the comment, is measured using Vader NLP [7]. Along with the above, keywords as well as the percentage of positivity of the comment is stored in the database, along with the ’pid’.

This helps in sending an effective report of usage, mindset and conversation around a particular content relevant to the brand.

RESULTS

From conducting the experimentation with the above pre-sented approach, it was found to have very positive and profitable impacts, for both, the platform and it’s client brands/corporations. The same can be proven by looking at the the facts about the platform. From the initiation of the project, the platform has currently seen 325,000+ users, with about a little over 150,000 new ones added from just the previous six months. The platform has as many as 20 major firms its regular clients. Some of the most notable ones who’ve been giving the most positive feedbacks about the services are MX Player, Amazon Prime Video, Alt. Balaji and Hotstar. The platform can be downloaded on major app. stores, on App Store and Play Store. All of the facts presented are from the time of writing this article and might be outdated at the time of reading the article.

CONCLUSION

The findings and results provided in this article provide strong arguments as to why the proposed approach is superior to currently widespread and conventionally used digital marketing strategies. It also shows how various subfields of Artificial Intelligence can help and aid in the marketing process and provide better guidance, value, and benefits to a corporation that utilizes the proposed approach.

ACKNOWLEDGEMENTS

We would like to thank MemeChat Pvt. Ltd. for sharing their data and allowing us to conduct the experiments and surveys which were heavily used in this article.

REFERENCES

1. Kavoura, A. (2014). Advertizing activities in social media and the creation of a community belonging in the digital era. Zeszyty Naukowe Małopolskiej Wyzszej˙ Szkoły Ekonomicznej w Tarnowie, (2 (25)), 97-106.
2. Chi, H. H. (2011). Interactive digital advertising vs. virtual brand community: Exploratory study of user motivation and social media marketing responses in Taiwan. Journal of Interactive Advertising, 12(1), 44- 61.
3. Truong, Y., & Simmons, G. (2010). Perceived intrusiveness in digital advertising: strategic marketing implications. Journal of strategic mar-keting, 18(3), 239-256.
4. Truong, Y., McColl, R., & Kitchen, P. (2010). Practitioners’ perceptions of advertising strategies for digital media. International Journal of Advertising, 29(5), 709-725.
5. Rice, S. V., Jenkins, F. R., & Nartker, T. A. (1995). The fourth annual test of OCR accuracy (Vol. 3). Technical Report 95.
6. Smith, R. W. (1987). The extraction and recognition of text from mul-timedia document images (Doctoral dissertation, University of Bristol).
7. Hutto, C. J., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media.
8. Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. arXiv preprint cs/0205028.
9. Prasad, D.S., Kabir, Z., Dash, A.K., Das, B.C.Effect of obesity on cardiometabolic risk factors in Asian Indians(2013) Journal of Cardiovascular Disease Research, 4 (2), pp. 116-122.
DOI: 10.1016/j.jcdr.2012.09.002