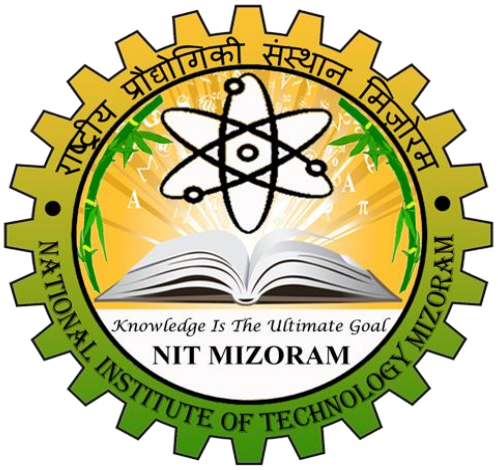
NATIONAL INSTITUTE OF TECHNOLOGY MIZORAM



ARTIFICIAL INTELLIGENCE ASSIGNMENT

Ruzina Haque Laskar	BT19CS021
Nitesh Raut	BT19CS016
Niraj Kumar	BT19CS031
Balajee Mishra	BT19CS033

Submitted to –

Dr. Debotosh Bhattacharjee
Professor ,
Department of Computer Science and Engineering .
Jadavpur University, Kolkata .

Personalize Social Media Accounts

Ruzina Haque Laskar , Nitesh Raut , Niraj Kumar , Balajee Mishra

Computer Science Department
National Institute Of Technology Mizoram

Abstract

The era of artificial intelligence (AI) has arrived. Companies all over the world are championing their latest progress with AI, machine learning and deep learning, even though most of it is far short of anything that could be described as a breakthrough. Promotion is one of the most effective ways to promote a business, and most people love promotions. Usually, these businesses announce their promo by uploading images to social media such as Instagram. However, most of the time these promo images are buried in the sea of other non-promotional images. It would be more practical if computers could be utilised to automatically look for images containing promotional offers. That is why this research is done to discuss creating a system that is able to tell whether an image contains information about a promotional offer or not automatically without human intervention using Optical Character Recognition (OCR) and Naive Bayes Algorithm as the classifier. Random Forest and K-Nearest Neighbour are also used as a comparison to the Naïve Bayes Algorithm.

INTRODUCTION

Social media is a widely spread phenomenon that has taken a large part of our lives. One of the major uses of social media is self-presentation (Kim, Kim, & Nam, 2010). Users can tweet and/or retweet on Twitter, build and manage friendships on Facebook, and/ or post photos with hashtags on Instagram. Each of these different services with different functions has its own strength in how users can share information about themselves. As of 2018, Instagram is one of the social media services that have grown rapidly. It is a photo-centric social media that is easy to use and has a convenient photo modification function.

Instagram is suitable for sharing information about what users see and how they want to be seen thanks to the diffusion of smartphones and improved camera performance.
(Kim, Kim, & Nam, 2010). Furthermore, Ferwerda, Schedl, and Tkalcic (2016a) suggested that Facebook users' personalities can be predicted by examining whether particular sections are disclosed in their profiles. Other studies have delved further into Social Network Services (SNS) posts in detail. Sorokowska et al. (2016) paid attention to selfies (self-taken photos) and found that extraversion predicted the frequency of selfie posting on SNS. Kern et al. (2014) showed that Facebook users' personalities were revealed through not only behaviour but also word choice and linguistic styles.

Concerning photo content, Liu, Preotiuc-Pietro, Samani, Moghaddam, and Ungar (2016) demonstrated that more positive emotions were displayed in agreeable and conscientious users' profile photos on Twitter, while users with a higher level of openness selected more aesthetic photos for their profiles. Additionally, users engage in interactions with other types of content like shares, following specific brands and reviews. it is possible to collect and analyse content concerning commercial purposes. The concern for privacy is substituted with the need to establish a social status or express beliefs regarding certain aspects of public life. Maintaining this status and reputation on specific platforms has become a marathon and social obligation in some circles. Social media has become the one-stop-shop for ideas, creativity, new products and lifestyle, is the corner piece to social and political moves and is the real estate for trendsetters. There is a shift in the communication model, users using the high bandwidth of mobile phones can now easily include photos, videos, gifs, maps, emoticons and other sorts of engaging content in conversations. For generation Z there isn't an unreachable location, unseen place or unknown product, the question is how to present it from an unconventional perspective and find new meanings. Getting valuable and actionable data from this controversial user-generated content is a marketing challenge.

These days, most people use social media for many things. Let us look at Instagram. More than 100 million photos are uploaded to Instagram every day. These photos belong to various categories, but sometimes people would like to see photos that belong to a specific category. So, it is necessary to be able to categorise these photos automatically. There have been previous studies where images were categorised based on the object that appears in those photos. However, the images used in those studies were images that don't contain any text at all. This is different from the case in social media where people also upload images that contain text. These texts might contain some information about the contents of the image and thus we can also take advantage of them to determine whether the image belongs to a specific category or not. In this study,

we are going to focus on detecting promotional offers that are uploaded by brands to their social media since these images of promotional offers will almost definitely contain some text.

Promotions are one of the many things that can make customers happy because they can save money to buy things that they want. Not only beneficial to the customers, but sales promotions are also beneficial to the businesses offering them as well. For example, when a business offers a discount for a specific item, customers tend to buy other things as well, this, in turn, increases the number of sales for the business. Promotions such as discounts could also enhance brand awareness, build customer loyalty, grow social media followers, improve reputation, as well as other things.

Promotions have so many benefits in our life as a customer, yet sometimes we miss them because we did not know that a promotion existed until the promotion period ended. We as customers will feel sad when we miss promotions which we are interested in. But the thing is, it is difficult to search and keep track of all promotions existing right now because as mentioned earlier, there are more than 100 million photos uploaded to Instagram every day, thus the information can easily get lost in the sea of photos of other categories. Then, there is also a possibility that the promotion we are interested in has ended when we just learned about its existence. To help solve this problem, we propose a system that can categorise whether an image contains promotions or not automatically without human intervention.
The scope of this article is to analyse social media behaviour and apply state of the art machine learning algorithms in order to find out how people interact and if particular patterns can lead to transactional marketing. The main focus will be on the concept of the hashtag and based on this notion we will do image analysis on Instagram using Google Cloud Vision API and verify if the post is a promotional offer or not automatically without human intervention using Optical Character Recognition (OCR) and Naive Bayes Algorithm as the classifier. Random Forest and K-Nearest Neighbour are also used as a comparison to the Naïve Bayes Algorithm.

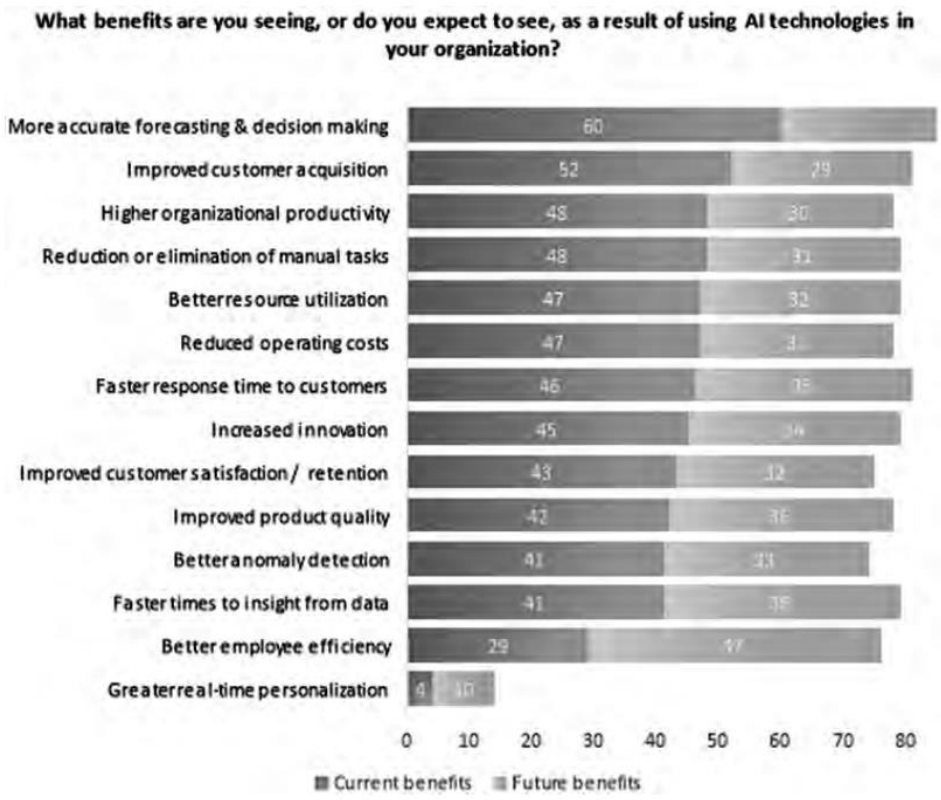


Figure 1: The benefits of AI

ALGORITHMS

1. **Support Vector Machine (SVM)** - This algorithm works by identifying a hyper-plane (refer to Fig. 2) that classifies the data points. Several different hyper-planes may be used to distinguish the two types of data points. SVM aims to find a plan with the maximum range, i.e., the maximum gap between the data points in both groups. Maximising the gap from the margins gives some clarification such that potential data points can be identified with better accuracy. Hyper-planes are boundaries for decision making and help to distinguish data points. Data points that fall on either side of the hyperplane can be assigned to different groups. Often,the hyper-plane dimension depends on the number of features. The hyper-plane is only a line until the number of input features exceeds 2. Whenever the number of features of the input is 3, then the hyper-plane would become a 2-D plane. It gets hard to picture when the number of features is greater than 3.

2. **K-nearest neighbours (KNN)** - KNN is a location-based approach used for classification and sometimes regression. This algorithm assumes that similarobjects occur next to each other and evaluates its k nearest neighbours for similarity. Training examples are multidimensional element space vectors, each with a class name. The algorithm's training process consists merely of holding the test sample feature vectors and class labels. k is a user-defined constant in the classification process, and an unlabeled vector (a question or checkpoint) is identified by assigning the mark that is most popular among the k training samples closest to that question point. The best choice IoT enabled convolutional neural networks for COVID-19 diagnosis and classification

3. **Naive Bayes** - Naive Bayes classifiers are a set of Bayes Theorem-based classification algorithms. It is called naive because all the features that are classified are assumed to be independent of each other, which is quite unrealistic in real life. The data is split into two components, the feature matrix and response vector. The feature matrix includes the whole data collection in the form of vectors(rows) where each row represents the relative variable type, whereas, in a response vector, each row represents an outcome class.

4. **Random Forest** - Decision trees are responsive to the particular data that is used to train them. If the training data is updated, the outcomes of the decision tree can be quite different. They are also computationally costly, bring a risk of overfitting and tend to find local optima because they can't go back after splitting. Random forests are used to fix these limitations. Random forest is an ensemble (multiple models combined) model technique in which multiple decision trees are trained together to produce one output. This merging of decision trees is termed as bagging.

5. **K-means** - It is one of the common techniques to cluster a dataset into several clusters. In K-means clustering, initially, the number of clusters (k) is initialized. Centroids are selected for N points randomly by tottering the dataset. Then the centroids are updated by taking the mean of all points within one cluster. The iterations must continue until the

clusters stop changing. For calculating the similarity, Euclidean distance or cosine similarity is generally used.

TOOLS

1. **Yotpo** - Another time-consuming chore that can now be managed owing to clever technologies like Yotpo is managing client reviews. Its moderating tool analyses consumer feedback and attitudes, giving you information into how to enhance your goods and services. It may also be used to increase sales by collecting important data from previous reviews and providing it to potential customers.

2. **AI-Writer** - AI isn't only a data-processing technology that delegates creative labour to people; it can also be used to create content. With simply a headline, AI-Writer can produce complete blog entries. Although the material isn't flawless, and the text output often needs some tweaking, this programme may nevertheless save content producers a lot of time when it comes to the labour of producing a new blog.

3. **Exceed.ai** - Exceed.ai is an artificial intelligence (AI)-powered sales tool that automates email communication and personalised conversation. This marketing automation technology tries to replicate human discussions by giving potential clients a natural experience, conversing with them, and learning their product needs before transferring the discussion to a real sales professional.

4. **NetBase** - Social media listening has been used by brands for decades, but NetBase takes it a step further by including machine learning and artificial intelligence. Its artificial intelligence (AI)-the powered system monitors millions of social media discussions in real-time to give real-time feedback on the audience's reaction to your newest brand news and changes. These insights may help marketers defend their brands, enhance crisis management, and increase campaign results.

5. **Node** - Node is an AI platform that uses data from people and organisations to anticipate how a company might engage its customers, workers, investors, and partners more effectively. The programme first establishes links between online elements such as people, goods, and companies, and then uses its algorithm to identify those who are most likely to convert or purchase. Marketing automation, management consulting, job applicant monitoring, and other applications may all benefit from these prognostic insights.

6. **Atomic Reach** -The AI technology from Atomic Reach is like having a content marketing assistant on staff. Its content switchover platform uses conversion optimisation and business analytics to help organisations convert their content into sales. While it can aid SEO, its findings are considerably more ROI-oriented, detecting what drives conversions and offering seamless improvements for blog headlines and tone.

7. **Seamless.ai** - Seamless.ai is a marketing platform that allows users to search and sift through big contact databases fast. Without the laborious process of list creation, contact research, data input, and other types of busywork, this sales intelligence application helps stack contact information (emails, mobile numbers, etc.) and then generate a list of prospects.

8. **Crayon** - Crayon uses artificial intelligence to automate competitive analysis. Its technology visualises competitive data in a single panel to keep advertisers up to date with the latest market news and developments, allowing them to spot patterns and make rapid, educated decisions.

9. **Unscreen** - Unscreen makes use of artificial intelligence to completely eliminate backgrounds from videos. Background free video used to necessitate complicated procedures, but with Unscreen, you can capture your film anywhere and then remove the background with ease. After you've eliminated the backdrop, Unscreen lets you replace it with a new static or video wallpaper.

METHODOLOGY

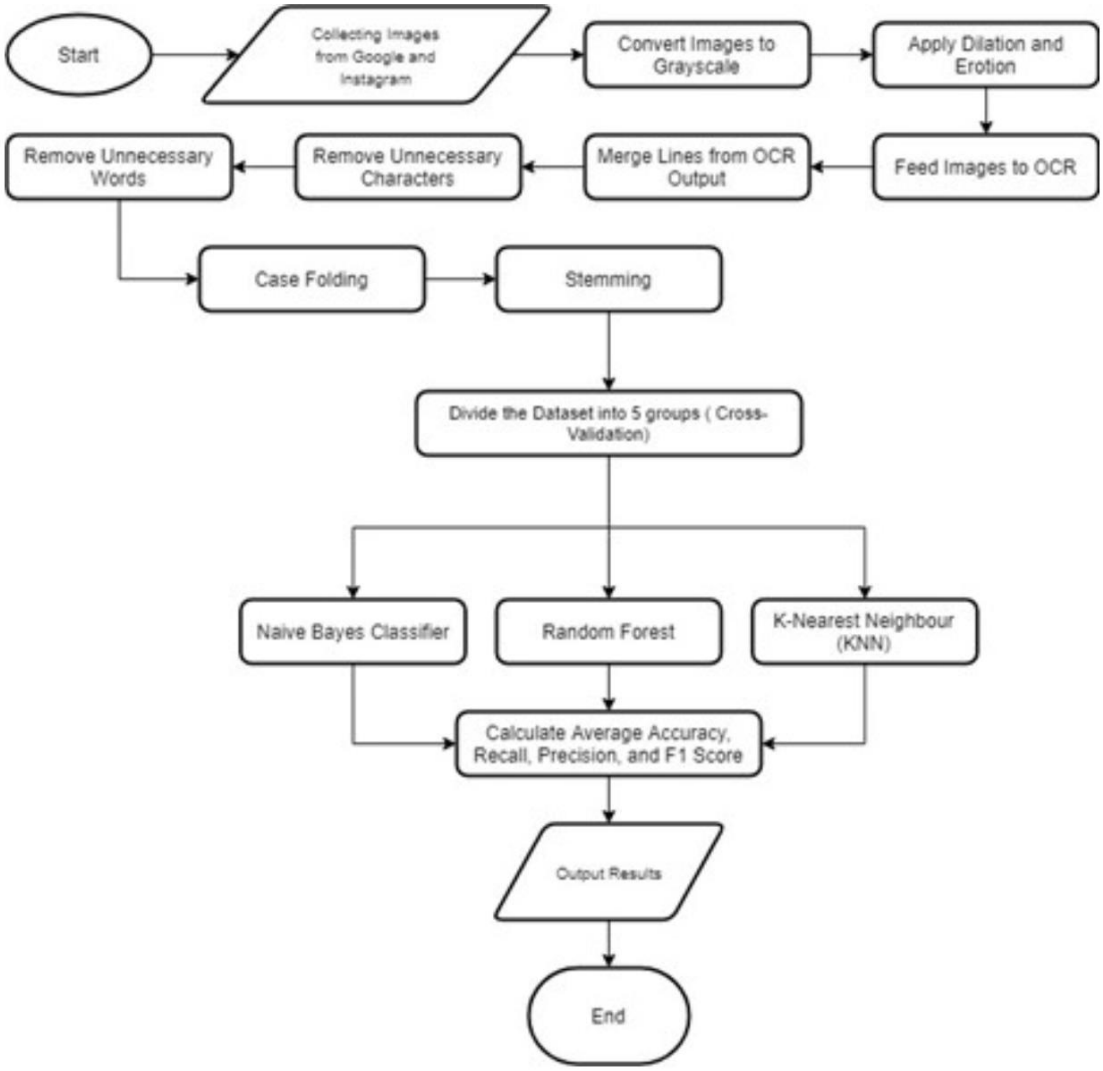
To achieve our goal which is to classify images to contain a promotional offer or not we are going to extract the text that is written on the images first and then classify the images with the extracted text instead of classifying the images directly using the individual pixel colour values that are contained within the images. This is because we humans use the text on the image as the main way to tell whether an image contains a promotional offer or not, whereas colours and objects that are contained within the images are just additional information. After the texts are extracted from the images, we need to preprocess it before they are used, which details can be seen below. Hereafter, we use classifier algorithms such as Naive Bayes, K-Nearest Neighbours (KNN), and Random Forest to classify whether the data contain promotion or not. Out of the algorithms we used, the Naive Bayes Classifier has the best result amongst all of them. Aside from the result, there are several reasons why Naive Bayes Classifier was chosen for this research, which are:

• Naive Bayes classifier is simple to implement.

• Naive Bayes classifier does not require much training data, in which we acknowledge that the dataset we use is lacking.

• Naive Bayes classifier is fast and can be used in real-time. This advantage may not be useful for now, but later this is important because the data used will be huge and will continue to grow over time.

• Naive Bayes classifier is not sensitive to irrelevant data. Even after preprocessing, there will still be words that have no connection to promotions at all.



A framework for extracting intelligence from Twitter data:

Collecting Twitter data (tweets and metadata) begins with identifying the topic of interest using a keyword(s) or hashtag(s), and requires the use of APIs (e.g., Streaming). This API method allows acquiring 1% of publicly available Twitter data. Twitter data is also available through data providers (e.g., GNIP, DataSift), also known as Twitter Firehoses, which can deliver 100% of Twitter data based on criteria. This is an ideal, but a very costly, option. Other social

media platforms also provide their API services. For example, Facebook offers Graph API.

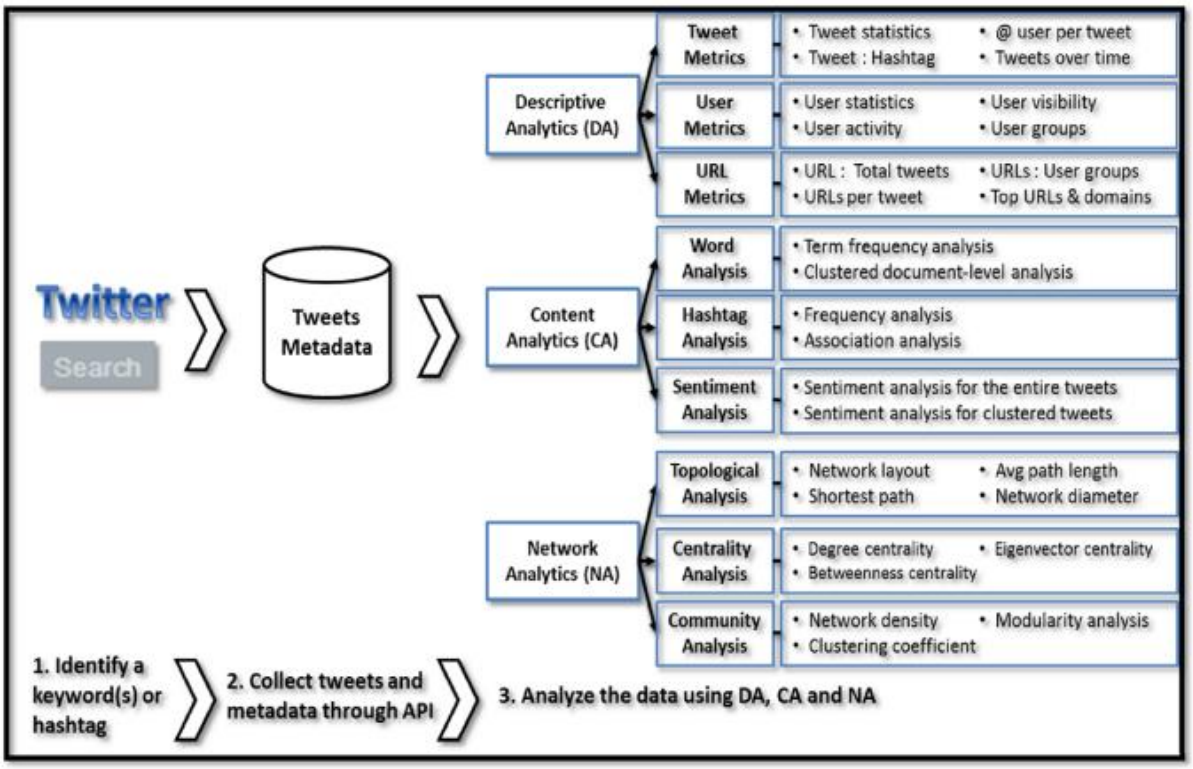


Fig. 1. A proposed framework of extracting intelligence from Twitter (Twitter Analytics).

A > Descriptive analytics (DA)

Twitter data contain a large amount of information, including tweets and metadata (e.g., user information). DA focuses on descriptive statistics, such as the number of tweets, distribution of different types of tweets, and the number of hashtags. Descriptive statistics are widely used in SCM research and practice. For example, researchers always report descriptive statistics about the survey they conducted. The difference lies in the number of metrics. While a small number of metrics (e.g., sample size, response rate, responder profile) are used for the survey data, the enriched nature of Twitter data enables intelligence extraction, using a large set of metrics regarding tweets, users, hashtags, URLs, etc (Bruns and Stieglitz, 2013) While only these three types of analyses are introduced, other descriptive analyses and metrics are certainly possible and should be used for different problems. However, the use of too many metrics is

likely to cause information overload and confusion, rather than intelligence. When using DA, practitioners and researchers should carefully consider a selective list of analyses and metrics according to the questions they are trying to address. Also, a change in analyses and metrics is expected when DA is used for other social media data (e.g., Facebook).

B > Content analytics (CA)

Social media data are primarily texts and thus "unstructured" in nature. Thus, it is necessary to use content analytics (CA), which refers to a broad set of natural language processing (NLP) and text mining methods, for extracting intelligence from Web 2.0 (Chau and Xu, 2012). A tweet's text is informal and composed of a shortlist of words, hashtags, URLs, and other information. Thus, careful consideration of text cleaning and processing is a prerequisite for intelligence gathering. Also, tweets (e.g., "#SupplyChain of cheap clothing stained with the blood of #Bangladeshi workers – HTTP:// nydn.us/181XT2M – via @nydailynews") contain not just information, but also opinions. Thus, advanced text mining techniques, such as sentiment analysis, are the key to extracting such opinions
Text mining and machine learning algorithms are important components of CA. Text mining transforms unstructured texts (or documents) into formatted data (or documents), using such techniques as tokenization, n-grams, stemming, and removing stop words (unnecessary words) (c.f., Weiss et al., 2005). Those transformed texts can be used for text summarization, keyword analysis, word frequency analysis, and text clustering, using machine learning algorithms, such as clustering and association analysis. While CA is found in supply chain research and practice (Georgi et al., 2010; Seuring and Gold, 2012; Vallet-Bellmunt et al., 2011), the approach has been manual or semi-manual, primarily through human interpretations. CA in TA relies on automatic text processing techniques and algorithms, due to the big data nature of Twitter data.

C > Network Analytics (NA)

Twitter users engage through @reply and retweet. As a result, it is possible to extract network information from Twitter data using the techniques and metrics in network theory, which is increasingly used in many academic disciplines (Burt et al., 2013), including SCM (Borgatti and Lin, 2009; Carter et al., 2007; Galaskiewicz, 2011; Kim et al., 2011). Nodes (e.g., Twitter users) and edges (e.g., relationships) are two basic terms in the theory. Network topology refers

to a layout of the nodes and the edges based on the information of replies and retweets on Twitter. This network visualisation uncovers patterns in interactions among users. Various network metrics (e.g., average path length) provide a detailed description of such a network. Using Twitter data, there could be two kinds of topological networks: friendship networks and @reply (or mention) networks. Friendship networks can be constructed based on the information of followers and following. Also, the conversation using @reply creates interpersonal relationships among Twitter users.
In addition, network theory offers centrality analysis, which uses node-level metrics, such as degree and betweenness centrality, revealing influential actors in the network. Degree centrality, a key metric, explains who has the most ties (or degrees) to others in the network (Wasserman and Faust, 2005). While degree centrality focuses on those nodes adjacent to a focal node, "betweenness centrality" includes distant paths of the focal node. While centrality analysis mainly focuses on individual nodes (or users in Twitter), community analysis explores network-level characteristics. For example, network density represents the portion of all possible connections between nodes, and, thus, it is a measure of network cohesion (Wasserman and Faust, 2005). Modularity is a measure of how strongly the network is divided into modules. Modularity analysis identifies specific communities from the network through visualization.One of the compulsory artificial intelligence fields in marketing is computer vision. At a basic level, computer vision is a technology that recognises patterns, and more importantly, understands them. Both science and engineering need to collaborate on computer vision. (Ali Borji, 2017).

A good match for the tremendous volume of user-generated data in social media enables the emphasising of machine learning algorithms. Furthermore, taking into consideration the ever-changing nature of the content is tiresome work to be labeled. For this operation, an unsupervised learning method like the K-means algorithm can be used to find commonalities and cluster the information for better managerial decisions. Unsupervised learning is a subset of algorithms in artificial intelligence that does not necessitate any prior training and can be utilised for clustering and segmenting information. More specifically using the K-means method one can divide any n number of assertions into K clusters, calculating the means between elements iteratively. Used commonly in computer science as clustering, this can be efficiently interpreted into marketing segmentation. A simplistic approach of the K-means clustering process described by Kyoung and Hyunchul (2008) is the division into four steps: a. an initial number of clusters K is established and placed randomly on the graph assuming they are centroids b. records are assigned to the nearest centroid forming k cluster c. taking into consideration the clusters resulting in step (2), a new centroid is calculated using the mean distance from each record in the

cluster. d. Iterate steps (2) and (3) until the centroid position stops shifting or conditions are satisfied.

Figure 1 shows that the segmentation of image tags is evident, for each of the 14 clusters was assigned a unique colour. Cluster 10 (in green) is differentiated away from the rest related to the cuisine and hospitality industry. Also, there is a slight overlap between cluster 12 (tourist, hill, mountain landscape, nature), 13 (beach, coast, sea, yacht, bae), and 9 (hairstyle, long hair, chin, forehead). This overlap might be determined by users who take pictures of themselves located at certain tourist attractions. An interesting fact observed was that the algorithm classifies differently the images from seaside, mountains and images from other attractions.

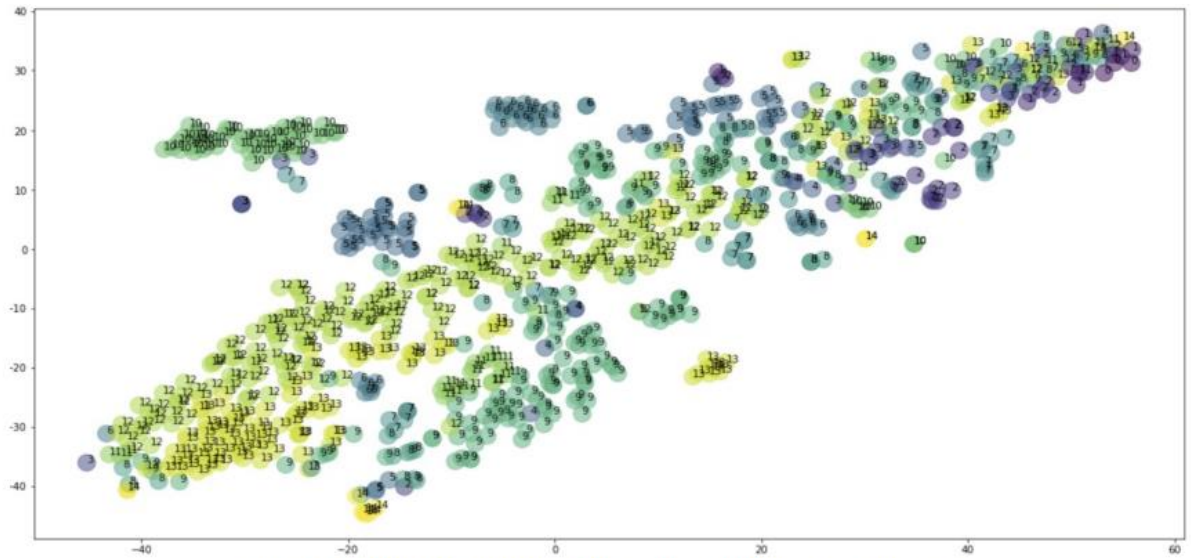


Fig. 1 Matplotlib output - clustering of instagram images

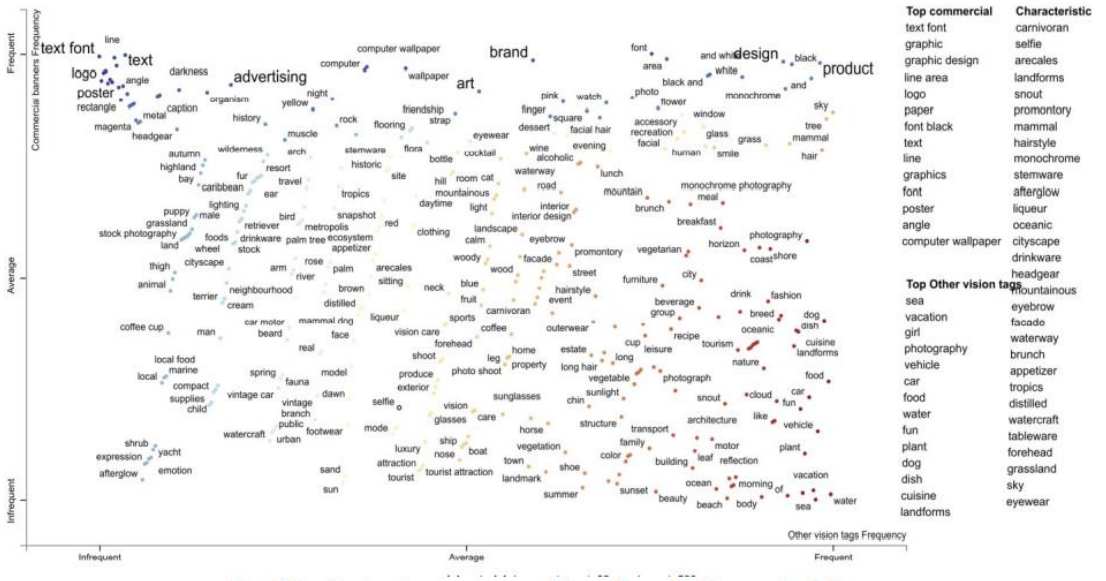


Fig. 2 Scattertext output - graphic and marketing materials

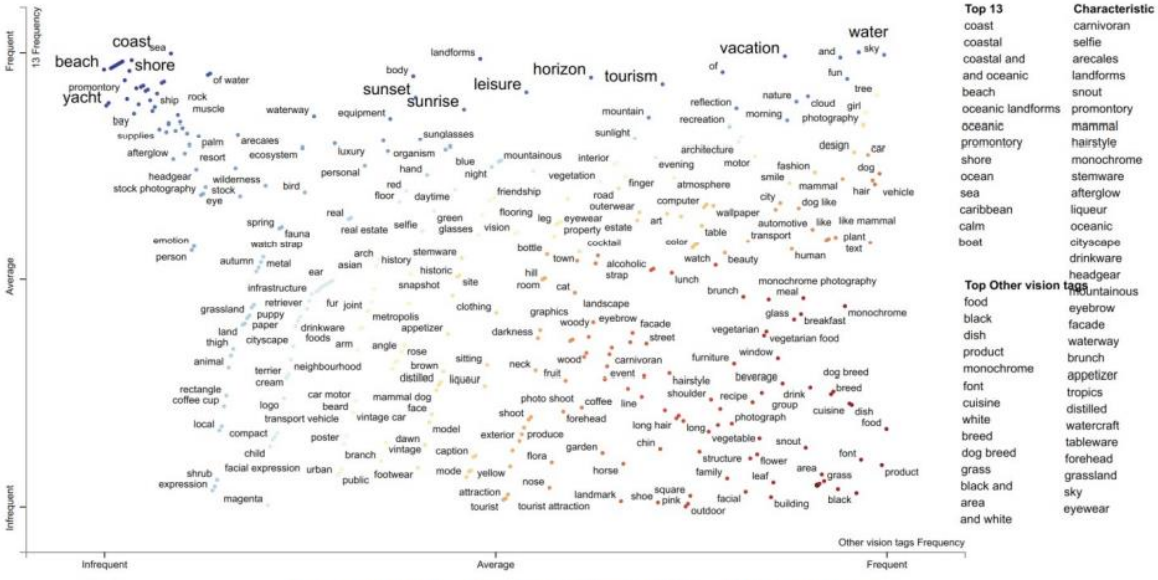


Fig. 3 Scattertext output - vacations, seaside images

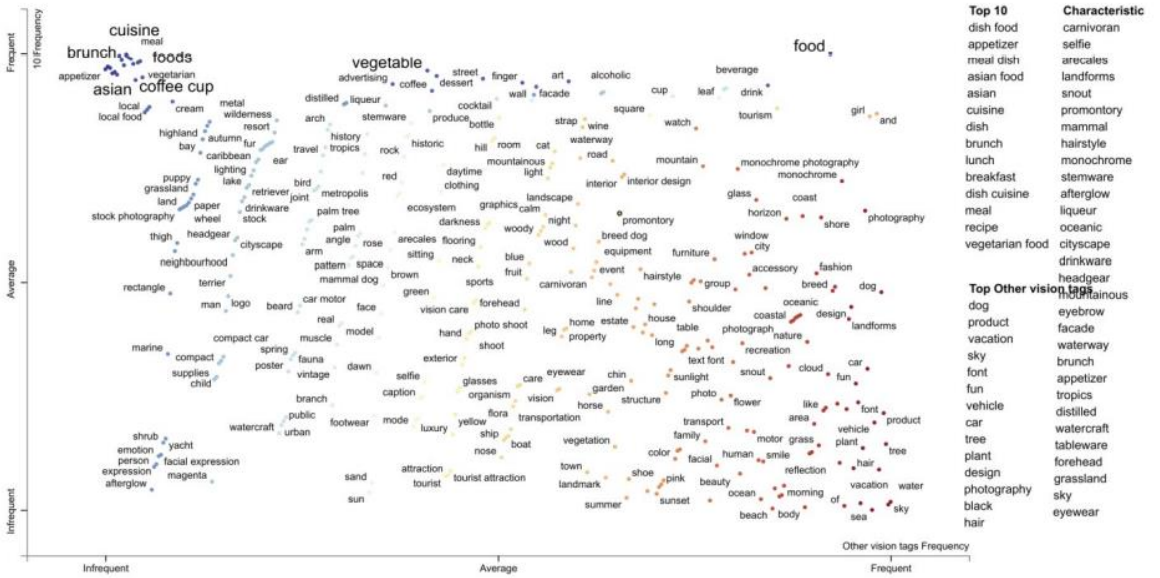


Fig. 4 Scattertext output - cuisine related labels with google vision

---

Discussion, implications and future research

Images are becoming the preferred medium of communication on social media platforms, understanding the information shared by the users and applying this knowledge into marketing strategies can represent a competitive advantage. This analysis has implications both in understanding user behaviour but also for better targeting the desired audience. By understanding how different keywords cluster together marketing managers can create cross-selling campaigns or promote the product including elements that usually are shown in images with the product.

Twitter ads have a feature that allows marketers to include specific keywords and it includes hashtags as well in order to target a specific audience. This option is not yet available on Instagram, but there is an option to target specific interests on Instagram and overlap this audience with a specific interest. Product presentation in e-commerce has implications on user purchases intentions, J. Park et all. demonstrated how the image size and better online visual product presentation created a pleasurable shopping experience. Using computer vision, analysing the content in the images can be taken one step further, allowing managers to identify cross-selling products, as visible in figure 5.

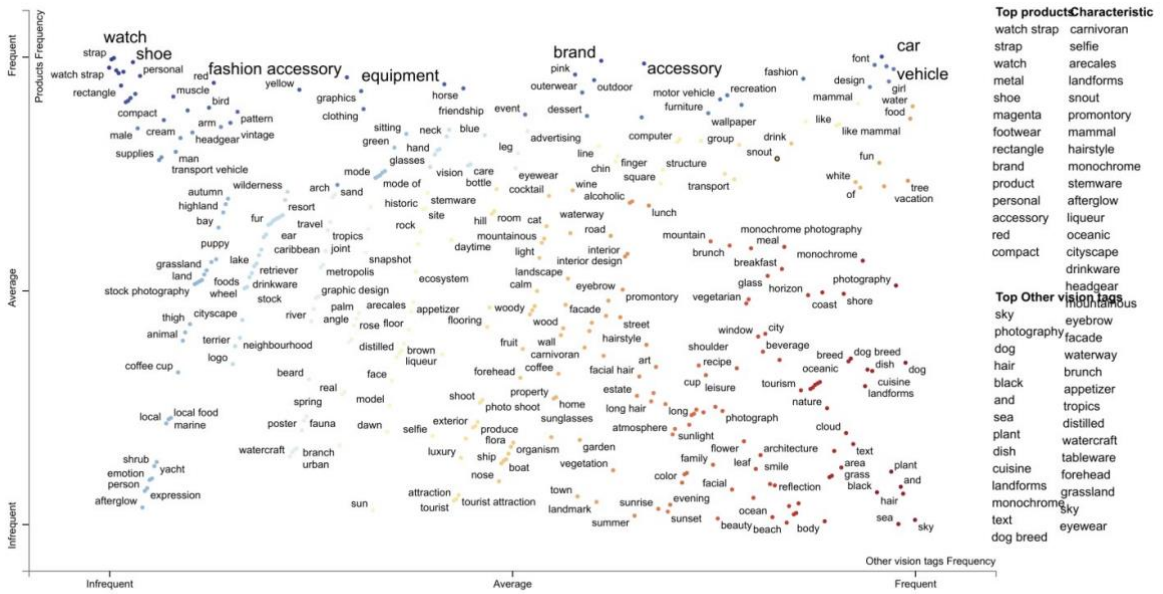


Fig. 5 Scattertext output - jewellery and accessory segmentation with K-means

Social media is an untapped source of data using the right analysis techniques can lead to substantial managerial implications. Making use of this chaotic stream of information is the domain of machine learning. Even if there are a multitude of options available on the internet, there is no substantial academic writing outlining how they apply to the field of marketing. Even if the plotting of clusters made the differentiation visible further empirical analysis is needed in order to confirm the findings. Future research can include the assessment of a statistical Chi-Square test of independence, to compare if there is a significant difference between the content of the hashtags (user-generated content) and the labelling done by a computer vision algorithm. In light of these insights, complex marketing applications can be developed taking into consideration dimensions like brands and products detected in images, geolocation in the metadata, sentiment analysis on description and comments

---

Conclusion

We notice the utility of artificial intelligence when the amount of data is enormous, which can leave the most experienced teams of data analysts and marketing researchers feeling disappointed. However, the processing of all this data is now easy using this technology. Moreover, this applies equally to other aspects of marketing and not just social media marketing. Artificial intelligence presents the next step of marketing campaigns; it allows generating personal information and using them for successful campaigns.

Companies now have the opportunity to use artificial intelligence technology to profile potential customers, analyse their behaviour, follow their habits, determine their motivations, etc. In order, to offer a product or service that meets their needs and expectations.
Social networks are a crucial playground for businesses, a personalised relationship with customers, but it does not prevent them from knowing that they are also highly saturated. However, the simple decision to use the strategic marketing tools of social media marketing is not enough; we must also rely on new techniques and technologies. That said, the word artificial intelligence can scare some people, but really, it excites avant-garde companies. Artificial intelligence technology is able to make marketing campaigns more personalised and smarter.

We can conclude that by using Optical Character Recognition (OCR) and Naive Bayes Algorithm it is possible to create a system that can automatically classify whether an image contains a promotional offer or not without human intervention. The Naïve Bayes classifier algorithm is the best choice to use for the system compared to Random Forest and K-Nearest Neighbour since it gives better results compared to the other two classifier algorithms with 94.31% accuracy and 94.33% precision on average. It shows a jump in accuracy compared to the previous research which has 75% of accuracy because of the improvement made on the preprocessing process such as applying stemming, modified character removal, and other minor processes. The system can then be implemented in the future for example as a part of an application which will constantly monitor for new images that are posted on social media and notify users once it detects an image that contains a promotional offer, this means that people won’t need to routinely monitor their social media anymore to discover new promotional offers as soon as possible. Some potential improvements that can be implemented in the future include providing a much larger dataset to train the model to increase the accuracy, doing further fine-tuning to the hyperparameters, and expanding the dataset to include images that are written in other foreign languages so that the system can also be beneficial for people of every part of world.

References

1. Kim, J. H., Kim, M. S., & Nam, Y. (2010). An analysis of self-construals, motivations, Facebook use, and user satisfaction. International Journal of Human-Computer Interaction, 26(11-12), 1077–1099.

2. Bruns, A., Stieglitz, S., 2013. Towards more systematic Twitter analysis: metrics for tweeting activities. Int. J. Soc. Res. Methodol. 16, 91–108.
3. Chau, M., Xu, J., 2012. Business intelligence in blogs: understanding consumer interactions and communities. MIS Q. 36, 1189–1216
4. Weiss, S., Indurkhya, N., Zhang, T., Damerau, F., 2005. Text Mining: Predictive Methods for Analysing Unstructured Information. Springer
5. Georgi, C., Darkow, I.-L., Kotzab, H., 2010. The intellectual foundation of the Journal of Business Logistics and its evolution between 1978 and 2007. J. Bus. Logist. 31, 63–109.
6. Burt, R., Kilduff, M., Tasselli, S., 2013. Social network analysis: foundations and frontiers on advantage. Annu. Rev. Psychol. 64, 527–547
7. Wasserman, S., Faust, K., 2005. Social Network Analysis: Methods and Applications. Cambridge University Press, New York