Solution Design for Detecting Duplicates on Classifieds (AVITO)

Figure: Image sourced from Pixabay.

1. Abstract

Creating duplicate postings on the classifieds is easy and most of the time exploited by the dubious users to spam and generate fake advertisements with the intention of theft. It also produces inconvenience for the people who are trying to run a business honestly on such websites. There is another problem of explicit content and listings that contains images or information that does not abide with the company's legal framework. So, it is important to bring down those listings and flag them to make a good experience for the end user. In this paper, the problem of ethical implications of duplicity is discussed in detail. The ethical discussion includes both on the frontend - where the user is interacting with the system directly and on the backend - where the human moderators are working hard to support a system which sometimes need external insight in labeling the listing as being duplicate. The paper proposes a framework solution using simpler yet more effective techniques from Image processing domain and Natural language processing domain, that are easier and effective when used together using Ensemble learning including a Neural Network. The paper concludes with the discussion on the prospective scalability offerings and the areas of improvements.

2. Defining Problem

Online marketplaces such as craigslist, olx.com and Avito all are market leaders in the classified spaces. These websites are basically buying and selling platforms or otherwise also called as classifieds that make it a breeze for the users to be a buyer, seller or may just be an advertiser. But since it is easy to create an advertisement – which contains both text and array of image data, it makes sense for some sellers to place the same advertisements with slight differences in text or may be in image – taken from different angle. Some advertisement like discussed above are copied/changed or replicated in wake of getting noticed amidst high volume listings and many competitive sellers might resort to this practice. But there are other bad players amidst this competition, that repost deceptive advertisements (altered in one or both text and images) with mal intention of sneaking money out of the sellers. So, it requires special attention to avoid the possible fraud.
Duplicate advertisement as such causes many problems such as that of fraudulent cases and even spam. To make sure that buyers can easily navigate to find what they want, without the worry of being enticed by a deceptive advertisement, there is always a need for developing a framework solution that can spot the duplicate advertisement to make it easier for the buyer to find and make their next purchase with an honest seller. In this way we are making sure that the buyer experience is good, and the sellers are not penalized for the actions of the bad actors in this classified space.

3. Schema Diagram and Data Filtration

This Data has been a part of a competition that was posted on Kaggle by Avito [12] (world's 3rd biggest classified market player). The data has been collected and sourced from Kaggle, however has been filtered down for the model due to limited processing available. Furthermore, to see if a listing is duplicate, we need to have a comparison to other listings in our data storage. And to be able to find that saving computation power is always a good idea. A simple filtration technique based on the data and domain knowledge can be done. I used simple filtration technique to reduce the size of the model based on categoryID. Since two items are more likely to be duplicate if they fall in the same categoryID.

itemInfo_train table contains several columns such as itemID, categoryID, title, description,images_array, attrsJSON, price, locationID, metroID, lat, lon. [12]

We are mainly interested in the categoryID as it is a simple heuristic that makes the result filtered before we apply the actual algorithm to the data.

Now that we have identified our candidates we can move to actual model and framework discussion. Note that there are other far more advanced techniques that deserve to be mentioned here such as Nearest Neighbor or K-Nearest Neighbor model. Those models are best to use if the data is humongous.

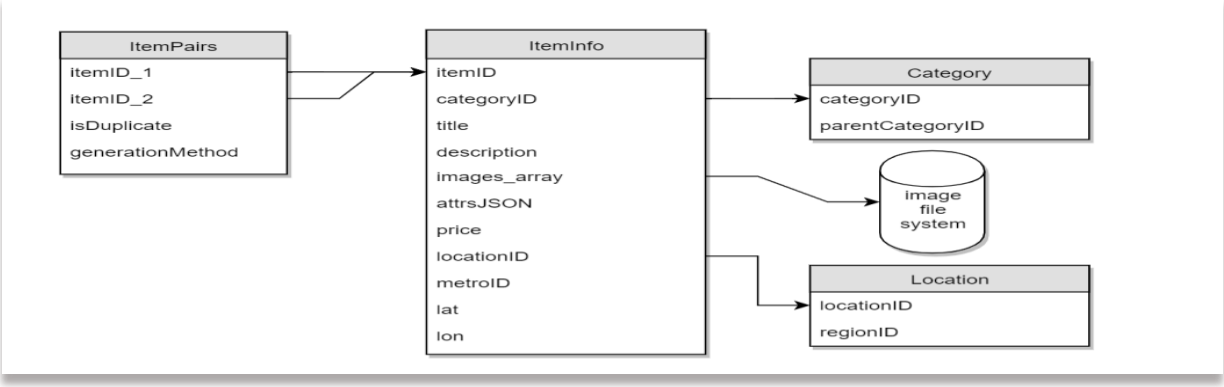
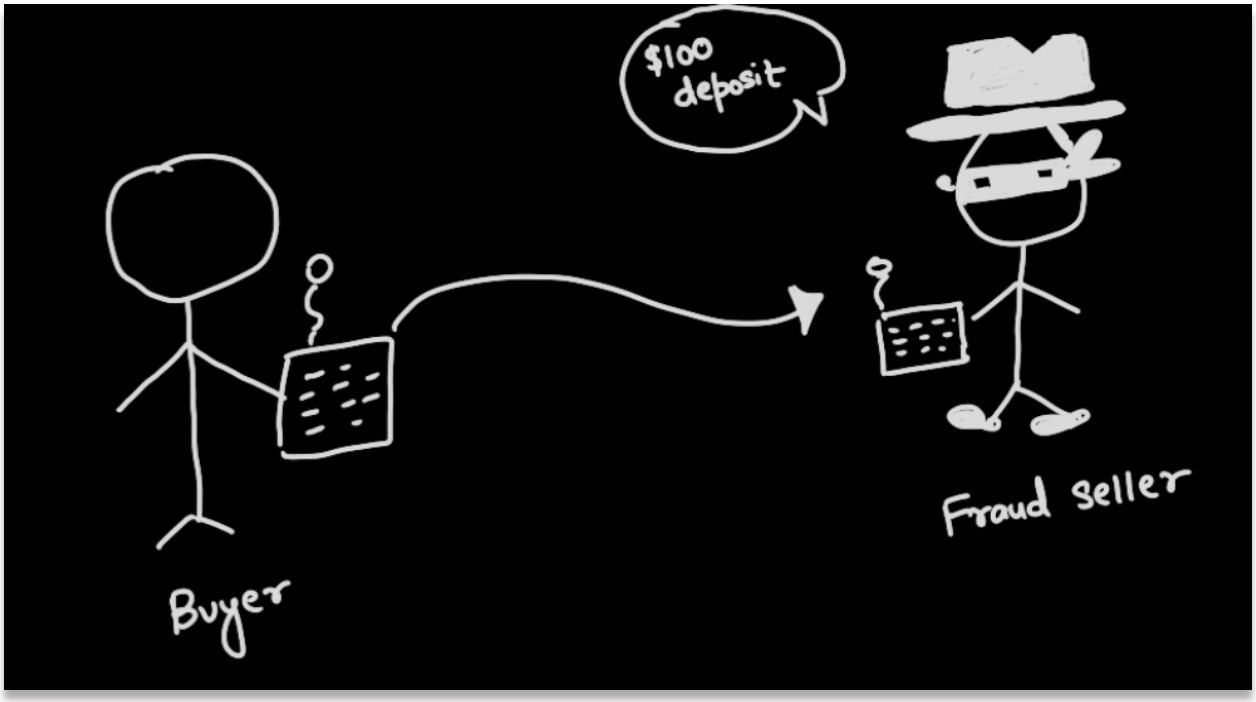


Figure: Sourced from: https://www.kaggle.com/c/avito-duplicate-ads-detection/data- The schema Diagram of the data.

4. Why is this important from Ethical point of view?

There are two problems from users creating multiple copies of the advertisements – one is spamming the platform and other is tricking the honest users into giving away their money causing a huge problem of problem. This places a mistrust on the platform by the honest users. Therefore, in my opinion the second problem is more serious and dire – fraudsters might trick users into giving away their money by replicating existing listings. And so, in this paper we are going to see an entire framework that can be designed to tackle the problem.



But this problem is not just limited to classifieds. When user is the one generating content such as on the social networking websites (e.g., Facebook and Twitter), online forums (e.g., Quora), classifieds (e.g., Amazon Marketplace, Craigslist, olx.com) – they all face the nuance of the user

generated content and left uncontrolled may cause problems more serious than even spam and fraud combined.
Many-a-times, dubious players try to sell something inappropriate; illegal may be such as guns, drugs, and other prohibited content. Such actions should be ethically handled by running a model that finds the duplicates and helps to tackle the problem to a much greater extent.

5. Understanding the flow that moderates what ads people can post.

When a user chooses to post a listing to the classified, it goes to the Machine learning based moderation system. The algorithm in that decides to accept, reject, or send it to the moderation queue where it then reaches the human moderators for more inputs. The human moderators (e.g., Amazon Mechanical Turk) then assign accept or reject score to it. Based on that the post for a listing goes live.
This is a scalable framework and parts of it being scalable have been described in the paper.

A typical illustration of how in organizations these flows work is as follows:

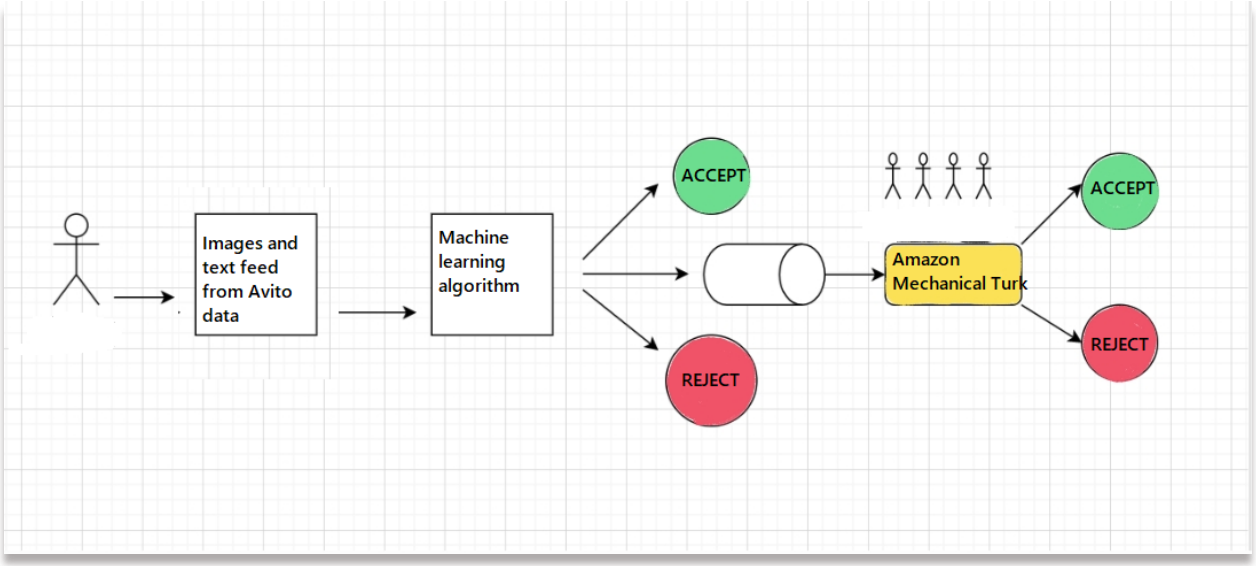


Figure: Illustration by Kritika Chugh.

The automated moderation system is typically a machine learning algorithm that have many complex features and components. We are going to see one such feature of filtering the post on a classified website.
Before moving forward with the framework, we shall discuss why is it more than necessary to have a good machine learning framework in place to filter out the content on classifieds or may be on social media for you, even when the human moderators are at play.

6. A case for Human moderators

The biggest tech giants in the world rely on AI to moderate the content. The domain might vary such as classifieds, social media, news services etc. but the goal is same. However, humans still

are a huge part of these content moderation workflow. The reason is that Artificial Intelligence (AI) even though has advanced a lot, but still cannot read between the lines. In other words, explaining AI the context is difficult and sometimes it makes mistakes. And for bigger brands to keep on maintaining consistency they outsource thousands of these manual content moderation jobs. The result is that they must sometimes go through explicit, violent and gore material and that affects their mental health. To top it all, they develop anxiety, fringe views and real fear of the world [8]. They are cheap labor working with 1/5th of the salary of an engineer and therefore one can safely say that they are even underpaid for the effects of work on their lives. And that builds an argument in favor of working on AI that can eliminate the need of such bad working environments.[9]

7. The Framework Design and Model Representation

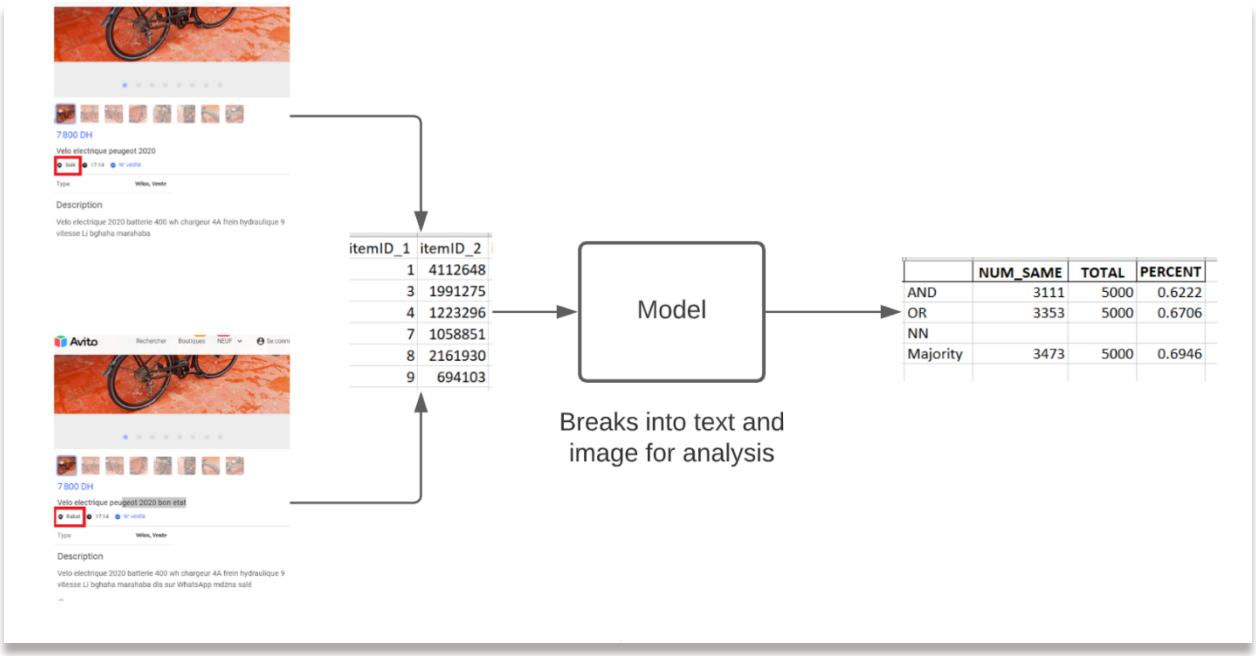


Figure: Pair of listing fed into the model that breaks it into Text and Image and performs comparison by yielding a score.

In the data sourced, we have **isDuplicate** column that tells us if the two listings are duplicate or not. This label is necessary for machine learning trainings. Typically, in organizations we first do mass manual annotations of this kind. It is important to have annotated data for supervised learning.

A	B	C	D	E	F	G	H
itemID_1	itemID_2	isDuplicate	categoryID_1	title_1	description_1	images_ar	attrsJSON
1	4112648	1	81	ÐŸÑ€Ð¾Ð´Ð°Ð¼ ÐšÐ°Ð¼ÐµÑ· 6520	ÐŸÑ€Ð¾Ð´Ð°Ð¼ ÐšÐ°Ð¼ÐµÑ· 6	1064094,	{"Ð'Ð¸Ð´
3	1991275	1	14	Yamaha r6	Ð'ÐµÑÑŒ Ð² Ñ‚ÑŽÐ¸Ð½Ð³Ð	11919573,	{"Ð'Ð¸Ð´
4	1223296	0	84	iPhone 3gs 8gb	Ð¢ÐµÐ»ÐµÑ„Ð¾Ð½ Ð² Ñ…Ð¾Ñ	14384831,	{"Ð'Ð¸Ð´
7	1058851	1	84	Xiaomi Mi4 3Ð³Ð± RAM + 16Ð³Ð± ROI	ÐžÑ‚Ð»Ð¸Ñ‡Ð½Ñ‹Ð¹		{"Ð'Ð¸Ð´
8	2161930	1	39	Ð›Ñ‹Ð¶Ð½Ñ‹Ðµ Ð±Ð¾Ñ‚Ð¸Ð¾Ð¿Ð¾Ð¸	Ð›Ñ‹Ð¶Ð½Ñ‹Ðµ Ð±Ð¾Ñ‚Ð¸Ð¾Ð¿Ð¾Ð¸	13718854,	{"Ð'Ð¸Ð´
9	694103	1	39	Ð¡Ð½Ð¾ÑƒÐ±Ð¾Ñ€Ð´ Ð±Ð¾Ñ‚Ð¸Ð¾Ð¿Ð¾Ð¸	ÑÐ½Ð¾ÑƒÐ±Ð¾Ñ€Ð´ Ð¸Ð‡Ð¾Ñ	12418395,	{"Ð'Ð¸Ð´
12	5637025	0	9	LADA Priora, 2015	ÐœÐ°Ñ‚Ð¸Ð²Ð° Ð½Ð¾Ð²Ð°Ñ Ð	1338189,	{"ÐœÐ°Ñ
12	5279740	0	9	LADA Priora, 2015	ÐœÐ°Ñ‚Ð¸Ð²Ð° Ð½Ð¾Ð²Ð°Ñ Ð	1338189,	{"ÐœÐ°Ñ
15	113701	0	32	Ð¢ÐµÐ»ÐµÐ²Ð¸Ð·Ð¾Ñ€	Ð¢ÐµÐ»ÐµÐ²Ð¸Ð·Ð¾Ñ€Ñ‹	11244051,	{"Ð'Ð¸Ð´
16	2467698	0	27	Ð¨ÑƒÐ±Ð° Ð¼ÐµÑ„Ñ‚Ð¾Ð¼Ð¾Ð²Ð°Ñ	ÐºÐ°Ñ‡ÐµÑÑ‚Ð²ÐµÐ½Ð½Ð°Ñ,	11762574,	{"Ð'Ð¸Ð´
19	4538621	0	88	ÐœÐ°ÑÐ°Ð¶Ð½Ð°Ñ Ð½Ð°ÐºÐ¸Ð´Ð¾	Ð Ð°ÑÐ¿Ñ€Ð¾Ð´Ð°Ð¶Ð¼ Ð¾ÑÑ	9722988	{"Ð'Ð¸Ð´
19	380482	1	88	ÐœÐ°ÑÐ°Ð¶Ð½Ð°Ñ Ð½Ð°ÐºÐ¸Ð´Ð¾	Ð Ð°ÑÐ¿Ñ€Ð¾Ð´Ð°Ð¶Ð¼ Ð¾ÑÑ	9722988	{"Ð'Ð¸Ð´

*Figure: **isDuplicate** column in the training data set acts as an important label for supervised learning*

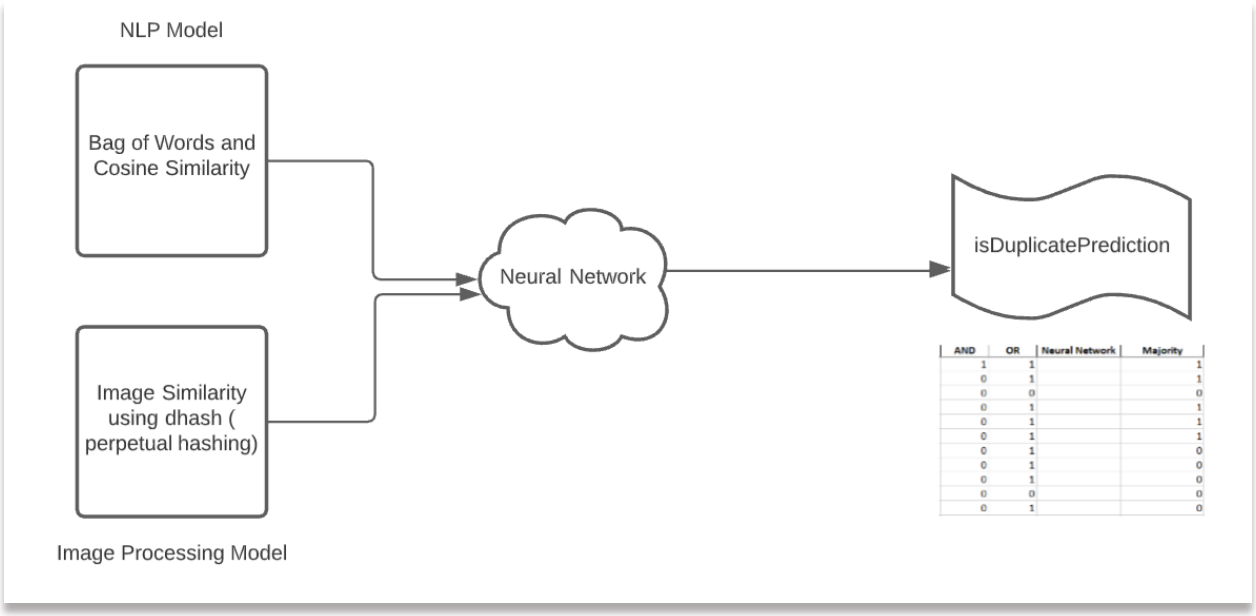


Figure: Model breakdown into 2 main functionalities: (i) Language Processing Model, and (ii) Image Analysis Model, that ensemble in a Neural Nework to predict the duplicacy of the listing.

Furthermore, the Framework has 2 features namely: Image feature and Text features which are discussed in the subsequent sections.

7.1 Text Features

On supplying the 2 listings to be compared, each listing is broken down into their text data and image data (discussed in next section). This is important to compare the text of the listing because many a times the advertiser/user just changes a little bit of description and title but keeps the same image and repost the listing under a slight change in data. This is how we can begin breaking down on the text data using a popular, yet effective technique called as “Bag of Words”. [6][4]

7.1.1 Bag of Words and Implementation

This method uses Word2Vec technique [4] to encode the text data to the vector form and then apply cosine similarity to find the angle between the two vectors. Smaller the angle between the vectors the more the similarity could be established between them.

Implementation:

- Text Preprocessing – In this step we prepared data by following:
  - Remove punctuation.
  - Removal of non-Ascii characters.
  - Removal of stop words.
  - Lemmatization of the filtered text.
- Feature Extraction – Using Word2Vec the words from the lemmatized sentence will now be represented as multidimensional array.
- Vector Similarity – Established using cosine similarity. Now that in Step 2 the word embeddings are available, we can compare 2 vectors. Cosine similarity gives cosine angle between the two.
- Decision made – If the degree of angle is less the cosine value would be high and vice versa. Lesser the angle more the text similarity. Furthermore, a threshold of 0.8 was added to make the decision into the model.[2]

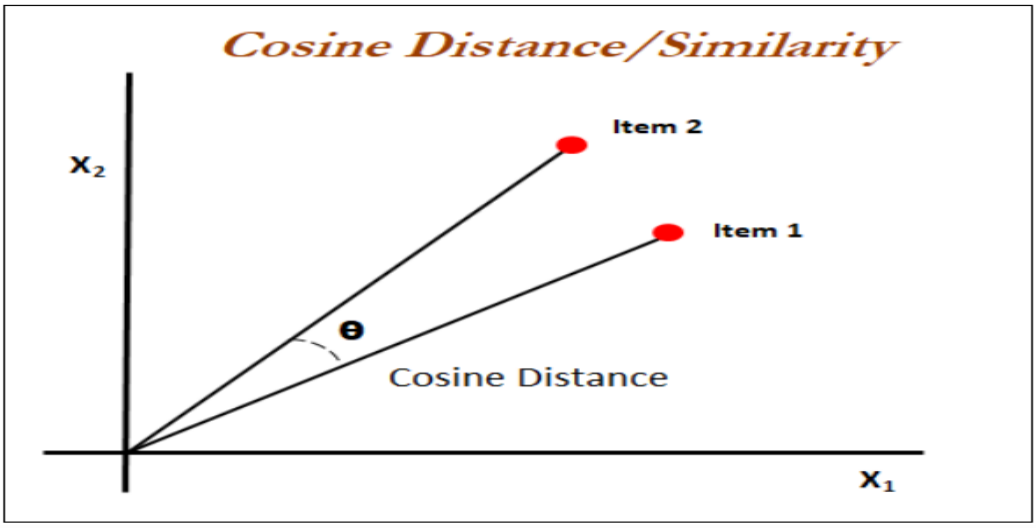


Figure: Sourced from O'Reilly

Image Similarity Predicted	Title Cosine Similarity	0.8	Desc Cosine Similarity	0.8
1	0.99999994		1	0.99999994	1
1	0.49999997	The threshold added to 0.8 for both title and Decription	0	0.816496611	1
0	0.666666627		0	0.522893608	0
0	0.942809045		1	0.99571085	1
1	0		0	0.801783681	1
1	0.288675129		0	0.876459777	1
0	1		1	0.445435375	0
0	1		1	0.491473168	0
1	0		0	0.523162723	0
0	0.49999997		0	0.69858253	0

Figure: Results from the NLP model with and without threshold.

7.2 Image Features

Image processing is the next half of the model for determining is two listings are similar or not. A listing comprises of both text and Image Array. Image Array in the sense that one listing/advertisement can have multiple images. Avito processes millions of images in a day and is therefore most difficult part of the duplicate detection. I used image hash feature to analyze the images [11]. Following is the duplicate advertisement found listed on the Avito website and it has different text and same image. The image difference was found to be '0' between the two listings – an indication of duplicity. The listings were only duplicate in the location that was fetched as the text and compared.

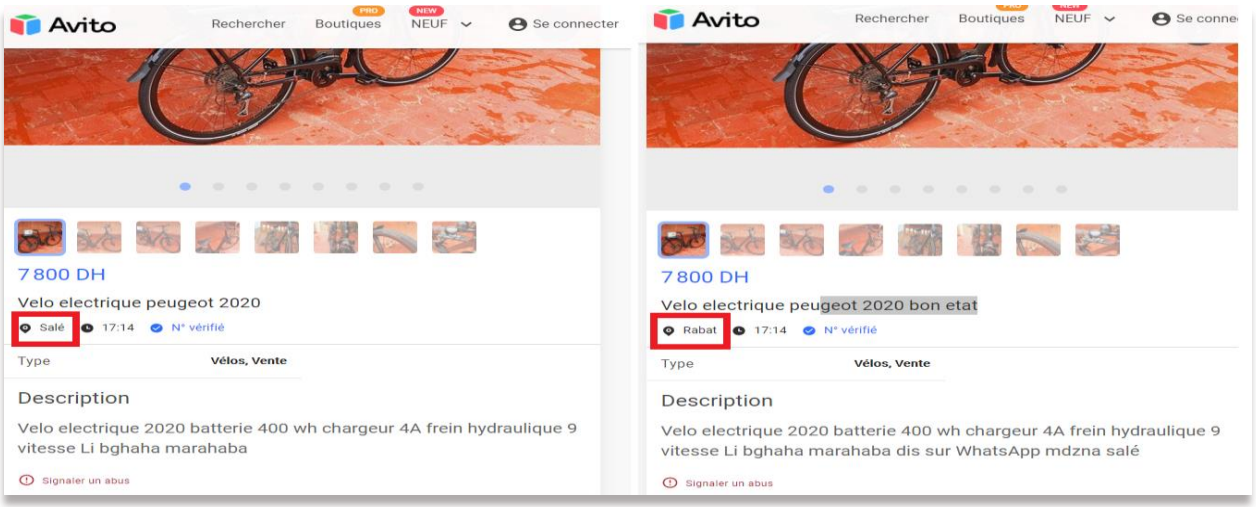


Figure: Result showing same image with hamming distance = 0, is an indication of duplicity in the listing.

7.2.1 Image Hashes and Hamming distance

Image hashing are of two kind and it is important to decide which one to use for the given problem. The ***Cryptographic Hash*** *[1]* and the ***Perceptual hash*** [5] are the two different kinds of hash that is available and its important to define the distinction between the two. The Image processing for this paper started with the experiments on MD5 but there was a problem. The

cryptographic hashes are designed to detect even the slightest change in the images. That did not come to the rescue as most of the duplicate listings are done with at least some changes to the images. These changes may be as simple as cropping the image or applying a filter.[10]
Moving on to the perpetual hash, we see that distance-hashing or d-hash is great as even crop and any change in resolution of the image will not affect the hashing value. And therefore, this works as a solution for this problem. [5]
To compare the images hamming distance has been used which gives us the difference in the number of bits in the d-hash computed for the image [11]. The smaller the hamming distance between the two d-hashes computed, the more similarity is established between the images and vice-versa. Additionally, it is worth noting that hamming distance is '0' for identical images.

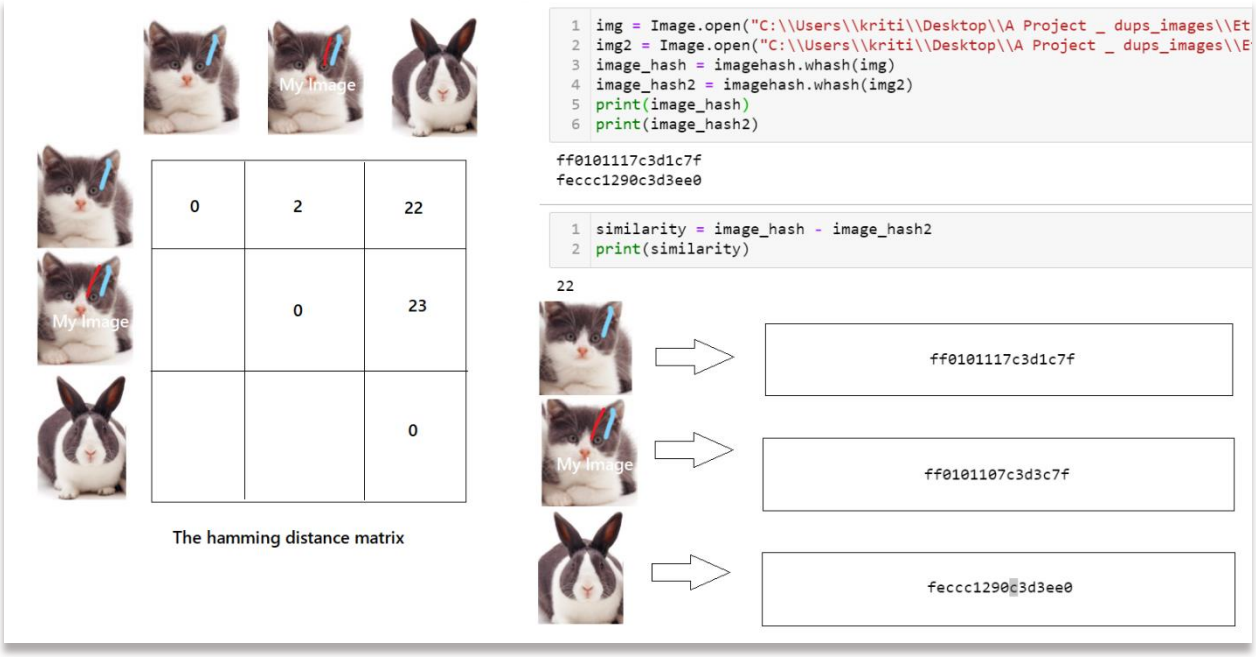
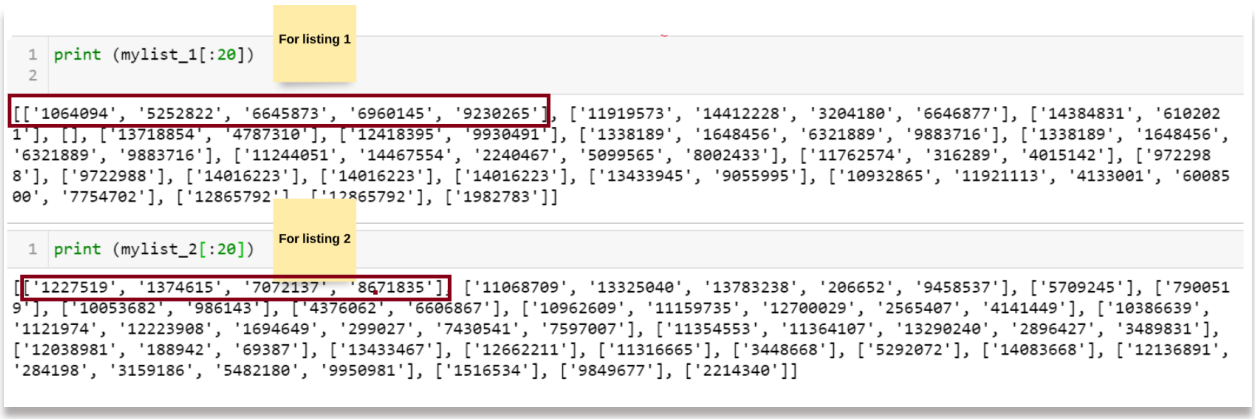


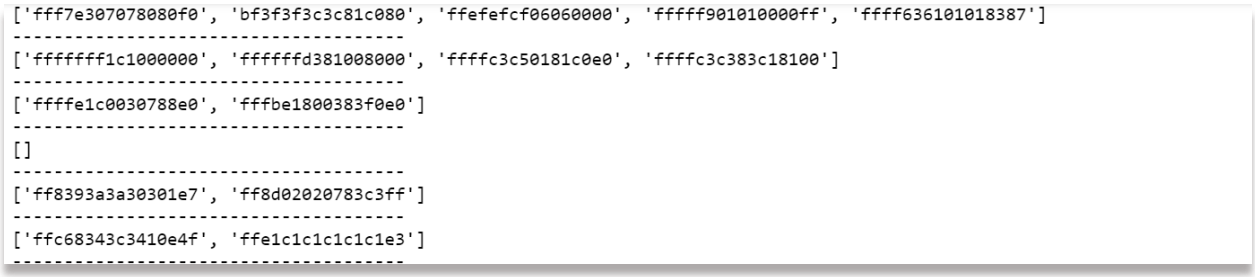
Figure: Three images with different hashes. The Hamming Distance between the top two hashes is closer than the Hamming distance to the third image.

7.2.2 Implementation

Loop over the data set of image array list for the 2 listings/advertisement to be compared. Each listing contains an array of images and once we have established the listings to be compared, we are going to perform a comparison between the images in the listings. Having an image in a listing is not a mandatory field and sometimes the data set contains some listing examples that are empty.



Grab each image and calculate the hash value using d-hash for each image. We are calculating hashing for two reasons. One it is lighter to store the hash values than the actual image and once comparison must be done the listing can readily be compared. This technique is not just faster but can also be scaled as we can always build a pipeline which calculates the hash value of each image that is uploaded in the system.



Build a rectangular grid out of two given one-dimensional arrays of images from each listing, representing the Cartesian indexing.

Now finally compute the hamming distance. For this Implementation I have constrained the algorithm to compute the number of images having a hamming distance less than or equal to 4. If even one image in the array list has a hamming distance less than or equal to 4, the algorithm will append that to the solution.[11]

This is done since during the training, as the findings based on the values of isDuplicate data suggests that even one duplicate image pair would yield a higher chance of the two listings being duplicate.

8. Ensemble into the Neural Network

The overall model is the sum of its parts – the text analysis and the image analysis. This is Ensemble learning [5], where we are utilizing two different models to realize a full solution with overall improved prediction on the duplicity of the classified listing/advertisement.

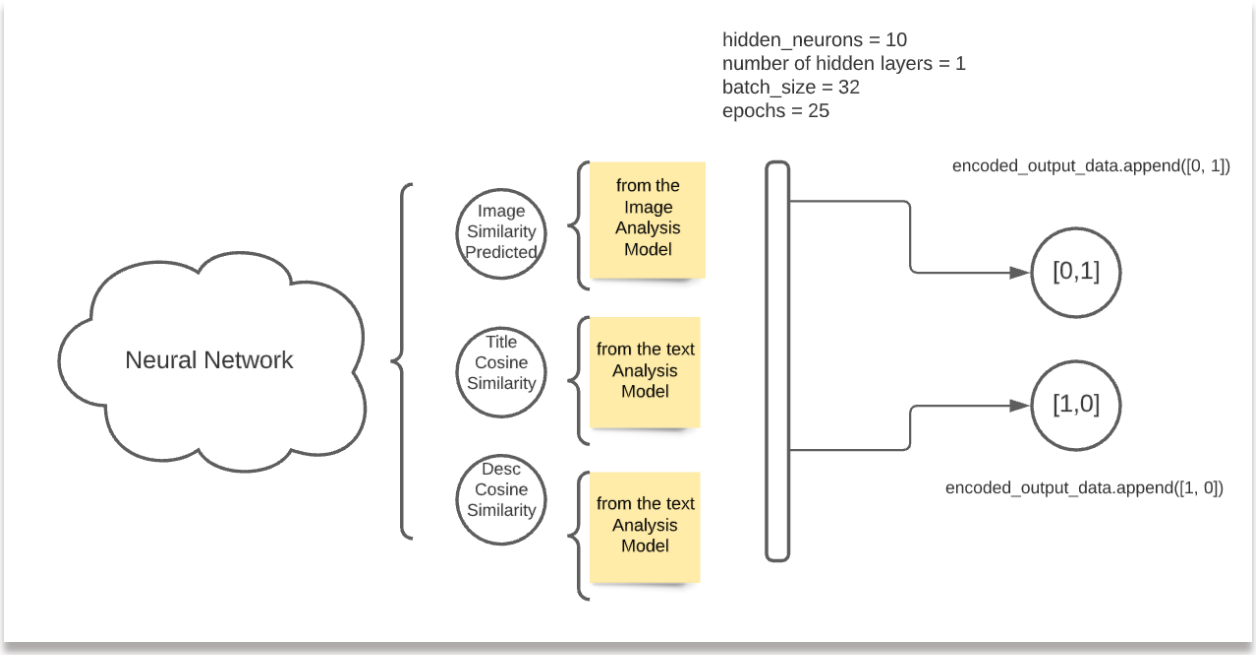


Figure: Ensemble training - The Neural Network Model Implementation

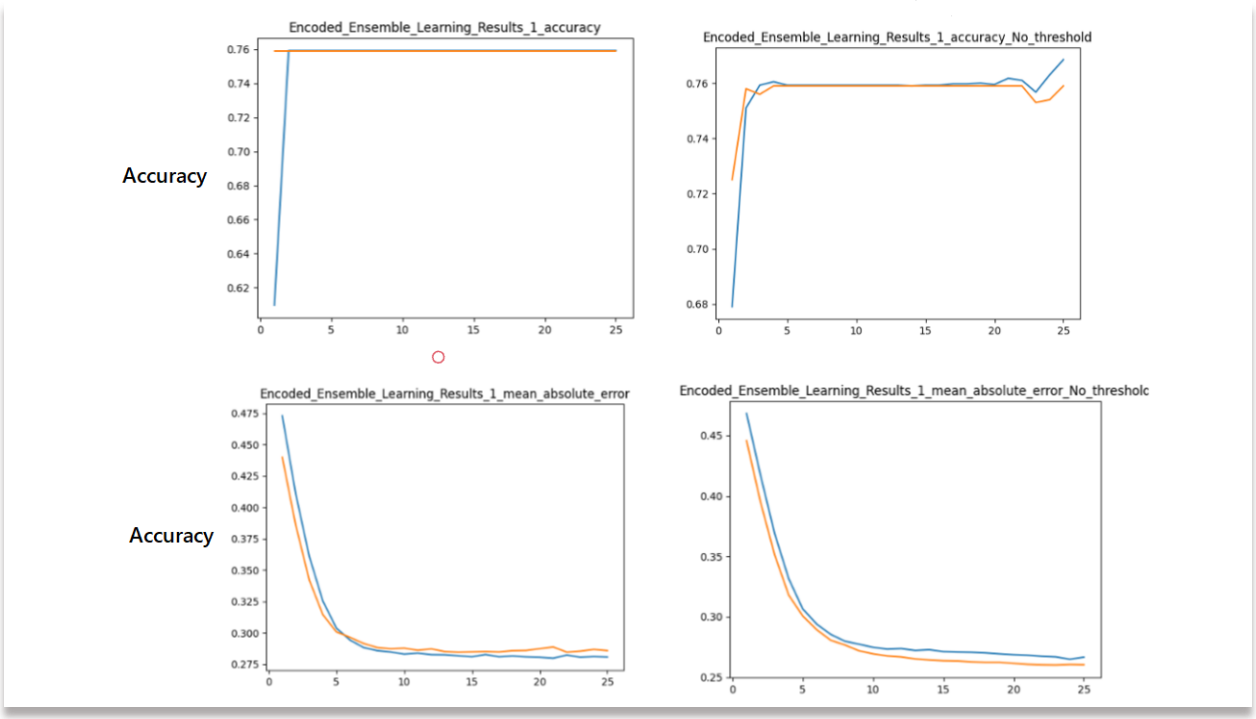


Figure: Results showing Accuracy and Absolute Error for a NN with 1 hidden layer. ——— represents training data and ——— represents test data.

Mostly we see that the accuracy is reaching as high as **~ 0.76** (plots) for both training and test data. Which is decent accuracy achieved compared to other models. Additionally, we can see

that although the model trained and performed on accuracy high the absolute error comes out to be low too. Both the results hold true with threshold and non-threshold values.
Furthermore, there is a slight overfitting on the accuracy when the model was tested with no threshold data.

The input test data has also been used with other simpler and underrated models such as AND, OR and Majority Models to compute the accuracy and to compare the values with that of the NN model. It turns out while NN beats the other models, but still Majority model being simpler and easier to explain seems to work decently for the problem. They seem important to be the part of this paper as a proof of explainable AI – a solution easier to understand by the humans. On the other hand, the NN model being tweaked to meet the needs of the current problem still is a kind of a black box to many. The point of including this summary is to compare and to make it easier for the user to comprehend what is being done and how well have we performed if we use NN instead of the other simpler models. This also is a part of the ethical problem, as many a times companies are required to explain the internal working of their algorithms for accountability and transparency reasons.

	NUM_SAME	TOTAL_RECORDS	PERCENT_ACCURACY
AND	3111	5000	0.6222
OR	3353	5000	0.6706
NN	759	1000	0.7590
Majority	3473	5000	0.6946

Figure: AND, OR and Majority Model accuracy comparison with the NN model

9. Conclusion

Having a moderation system deployed on classified platforms ensures high quality content and discourages the fraudulent advertisement to cause any cases of fraud from happening.
In many cases such moderation frameworks work in conjunction with image detection and language detection mechanisms to bring down the content which is illegal or violates the company policies.
The Model framework consists of “Image Similarity” using perpetual hashing [5] and hamming distance and Natural language processing model utilizing “Bag of words” approach that encodes text into vectors and then finds the cosine similarity between them.
Using hashing is the easiest and the fastest way to bring images into the framework code and perform comparison on them as they are lighter to save in the disk or the databases.
The first half of the paper discussed how simple understanding of the domain can help us narrow down the candidates/listings/advertisements we want to work with to eradicate duplication.
The second half talks about the machine learning approach (using image similarity processing and NLP concepts). The two models both NLP and image processing gives us classifications which then are ensembled to give us the final output. The output tells us if a particular advertisement was indeed a duplicate or not.

The final section contains the comparison of the 4 Ensemble models (AND, OR, Majority, Neural Network with one hidden layer) and why the simple approaches such as AND, OR and Majority are important from the perspective of explainable AI.
There are several areas this project has shown me and sparks an interest in the scalability as its future work. Scalability is one of the biggest issues when it comes to expanding such ideas on a large scale. For example: the matrix created for calculating the hamming distance has O(n^2) complexity and these are huge when you scale the system to say 1000,000 images. In such cases other searching mechanisms would fail.
The model could also be scaled in other dimension where in we seek the help of third-party tools and create a pipeline which calculates hashes and puts it to the storage every time an image is uploaded as a part of the listing/advertisement on these classified websites. This would make the queries faster to perform and overall performance would be high always. It also brings modularity to the framework as it removes the management of the hashing problem from the code.
There are other hashing techniques too that exist which can be utilized in conjunction with d-hash to improve the accuracy even further. This is important as listings are more likely to be the same if images are the same.
Even though this paper has discussed the ethical implications of the human moderators [8][9] in such filtration mechanisms, we should still note that the dependency will not go away immediately. There is no panacea for this problem. So, in essence the realistic approach would be to improve the algorithm and making sure to save the decisions made by the human moderators during the filtration technique and that can be used to improve the overall performance of the model in the future.

10.References

[1] Rupali Roy, towardsdatascience.com, https://towardsdatascience.com/hiding-data-in-an-image-image-steganography-using-python-e491b68b1372 , “Hiding images using Steganography and RGB”, Last Modified: May 7, 2020
[2] Paul Minogue, https://paulminogue.com/index.php/2019/09/29/introduction-to-cosine-similarity/ , “Cosine Similarity and Sentence Vectorization”
[3] Thushan Ganegedara, https://towardsdatascience.com/light-on-math-ml-intuitive-guide-to-understanding-glove-embeddings-b13b4f19c010 , “GloVe Model and Embedding Model”, May 5, 2019
[4] “Wikipedia.com” https://en.wikipedia.org/wiki/Word2vec , Word2VecModel.
[5] “Wikipedia.com” https://en.wikipedia.org/wiki/Perceptual_hashing#:~:text=Perceptual%20hash%20functions%20are%20analogous,drastic%20change%20in%20output%20value.
[6] NLP Python by “Edward Lopper”- http://www.nltk.org/book/
[7] Aditya Oke, Medium.com, https://medium.com/pytorch/image-similarity-search-in-pytorch-1a744cf3469 , “understanding of how image similarity can be done using PyTorch”
[8] https://www.researchgate.net/publication/335923872_Re-humanizing_the_platform_Content_moderators_and_the_logic_of_care - “Ethical standpoint discussion”

[9] Casey Newton, The Verge, https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona "Ethical discussion on what human moderators go through"
[10] Adrian Rosebrock https://www.pyimagesearch.com/2017/11/27/image-hashing-opencv-python/
[11] Kim, Jun-Seob, Jung, Wookhyun, Kim, Sangwon, Lee, Shinho, Kim, Eui, Evaluation of Image Similarity Algorithms for Malware Fake-Icon Detection https://www.researchgate.net/publication/347806326_Evaluation_of_Image_Similarity_Algorithms_for_Malware_Fake-Icon_Detection
[12] "Kaggle.com | Avito.com" https://www.kaggle.com/c/avito-demand-prediction