A Comparative Analysis of Online Reputation Monitoring

- **Sarvottam Kumar Modi**
- **18CH10073**
- **Sarvottamkumar.iitkgp@gmail.com**

Overview

Initially, I have written summary on 12 research papers to understand the topic I was going to work on and other information that will be required for the project. This also gave me an idea about how the research papers are written and extract useful information from them. Data mining of 90,000+ tweets and its related information from tweet IDs in the Replab Dataset 2013 were done by accessing the Twitter API using the Tweepy library of python**.** Exploratory Data Analysis of tweets like text statistics, N-gram exploration, WordCloud, and Named Entity Recognition was done to understand the pattern of data and other features of the tweets. WordCloud of hashtags and overall text gave us an insight into the popular topics on which discussion was going in Twitter. Named Entity Recognition was done to extract information about the different brands of car and their model. It also gave us information about the different places and timelines. Sentiment Analysis of tweets using the libraries - Afinn, Textblob, and NLTK's Vader Sentiment Analyzer gave us information about positivity, negativity and neutrality of tweets. Topic Modelling was done using Genism's LDA module and determining the number of topics & its distribution over tweets. PyLDAvis library was also used to visualize the topics and the top most occurring words of different topics. Brand Popularity of entities from tweets using cosine similarity and number of retweets & likes.

Writing small reports on research papers

In this task, I needed to write small reports on 12 research papers gave by the mentor. It consisted of sub-topics like aim, idea, definition, and some crucial points about the report. Sometimes it also included a brief description of the topic and different approaches used in the paper for the technique described in it. These reports described what a research paper contains as a whole and different ideas & terms used in it. The research papers I read gave me a mere idea on the topic of online reputation management, brand-popularity and sentiment analysis.

Here is glimpse of some of the reports which I have summarized :-

1. Category Specific Post Popularity Prediction https://doi.org/10.1007/978-3-319-73603-7_48
Aim: To predict popularity of a brand by specific category posts
Idea: Predicting popularity of a post categorically using its visual and textual content
Important points :-

- Previous works focus on predicting the popularity without considering the category type of the post.
- Predicting popularity of posts has become interesting for marketing and business, political and economic sciences and decision-making strategies of campaigns targeting social media crowds.
- Predicting post popularity is important for the self-evolution of social media.
- Instagram has a strong emphasis on self-expression by images with a description through captions and hashtags and is easy to crawl.

- For the purpose of this paper a new dataset was created from scratch, containing posts crawled from Instagram.

In this paper, they proposed a multimodal framework for post popularity prediction especially when action, scene, people, and animal appear in the users' posts. They performed several experiments on a collection of 65K posts crawled from Instagram. Most of these posts were related to activities, such as dancing, landscapes, such as beach, people, such as a selfie, and animals such as pets.

2. A Spatio-Temporal Category Representation for Brand Popularity Prediction
https://doi.org/10.1145/3078971.3078998
Aim: To predict popularity of brands based on various social media posts
Idea: category based popularity prediction by integrating the *spatio-temporal* dimension
Methods and their explanations:-

- Existing instance-based popularity prediction methods focus on popularity of images, text, and individual posts.
- Some factors of popularity such as user activeness and image prevalence variability can be viewed as time-sensitive, means that for most accurate analysis of popularity one needs to address *temporal dynamics*.
- automated systems have been developed to extract important information from different types of user generated content using *domain agnostic methods*.
- Previous papers explain how business profit by using social media as a source of information for business intelligence as well as an execution platform for product design and innovation, relationship management, and marketing.
- This paper incorporates the *temporal dimension* in the brand representation to account for changes over time in popularity that are not solely reflected by the content.
- top brands on Instagram are seeing a per-follower engagement (i.e. likes per follower) rate of 4.21 which is 58 times higher than Facebook and 120 times higher than twitter resulting in study of datasets built by Instagram.

This paper uses a new method to analyze social media posts generated by various brands during a specific period of time. It proposes a *spatio-temporal* category representation for brand popularity prediction. This paper studies the behavior of this method by performing four experiments on a collection of brand posts crawled from Instagram with 150,000 posts related to 430 active brands.

3. Customer engagement, buyer-seller relationships, and social media
https://www.emerald.com/insight/content/doi/10.1108/00251741211203551/full/html
Aim: To develop a framework and examine practitioner views on customer engagement.
Ideas: develop a customer engagement matrix on the degree of relational exchange and emotional bonds between buyer and seller.
The paper include following ideas:-

- a better understanding of the concept is essential to develop strategies for customer engagement.
- enhance understanding of customer engagement by:-
    - examining practitioner views of customer engagement,
    - linking it to the marketing concept, market orientation,
    - relationship marketing, modeling the customer engagement cycle,
    - developing a customer engagement matrix

- Four types of relationships emerge with customers:
    - transactional customers, delighted customers, loyal customers, fans.
- A mix of digital and nondigital technologies can be employed to facilitate customers' transition through the stages in the customer engagement cycle.
- It provides the foundation for strategies to better satisfy customers using Web 2.0 tools like social media.
- As exchange relationships evolve from market-mediated exchange to organization-mediated exchange, more intimate and enduring relational exchanges take place.

The paper develops a model of the customer engagement cycle. The advent of the internet and in particular the interactive features of Web 2.0 in recent years have led to an explosion of interest in customer engagement. Companies are striving to create a high level of customer engagement defined as "an intimate long-term relationship with the customer". This paper formulates a customer engagement cycle with connection, interaction, satisfaction, retention, commitment, advocacy, and engagement as stages in the cycle.

6. Evaluation of unsupervised emotion models to textual affect recognition
https://www.aclweb.org/anthology/W10-0208
Aim: To evaluate new Unsupervised Emotion Models for detecting emotions in text.
Idea: To study categorical model and dimensional model for the recognition of four affective states : Anger, Fear, Joy, and Sadness.

Different models used :-

- In the first model, WordNet-Affect is used as a linguistic lexical resource and three dimensionality reduction techniques are evaluated: *Latent Semantic Analysis* (LSA), *Probabilistic Latent Semantic Analysis* (PLSA), and *Non-Negative Matrix Factorization* (NMF).
- In the second model, *Affective Norm for English Words* (ANEW), a normative database with affective terms, is employed.
- Experiments show that a categorical model using NMF results in better performances for SemEval and fairy tales, whereas a dimensional model performs better with ISEAR.

Learning techniques used :-

- Supervised learning techniques have the disadvantage that large annotated data-sets are required for training.
- And, the process of the annotation is very time consuming and expensive.
- For this reason, unsupervised methods are normally preferred in the realm of *Natural Language Processing* (NLP) and emotions.

Previous techniques used :-

- (Strapparava and Mihal-cea 2008) - describe the comparison between a supervised (*Naïve Bayes*) and an unsupervised (*Latent Semantic Analysis* - LSA) method for recognizing six basic emotions.
- (D'Mello, Craig et al. 2008) - used LSA but for detecting utterance types and affect in students' dialogue within Autotutor.
- (D'Mello, Graesser et al. 2007) - proposed five categories for describing the effect states in student-system dialogue.
- (Kort, Reilly et al. 2001) - proposed (but provided no empirical evidence) a model that combines two emotion models, placing categories in a *valence-arousal* plane
- (Aman and Szpakowicz 2007) - studied how to identify emotion categories as well as emotion intensity.

This is the link of the doc given for all the reports of research papers I have wrote :-
https://docs.google.com/document/d/1QR_-7k3k6T1CwYjQxQDs3hvulVoaXhKOp4dSQjmA-Ks/edit?usp=sharing

All of the research papers which I have summarized are titled as follows :-

1. Category Specific Post Popularity Prediction
2. A Spatio-Temporal Category Representation for Brand Popularity Prediction
3. Customer engagement, buyer-seller relationships, and social media
4. Evaluation of unsupervised emotion models to textual affect recognition
5. A rational analysis of marketing strategies
6. Automatic generation of entity-oriented summaries for reputation management
7. Integrating learned & explicit document feature for reputation monitoring in social media
8. Multimodal Popularity Prediction of Brand-related Social Media posts
9. A mixed-methods approach for exploring patterns of communication-related to business scandals on Twitter
10. Comparison of social media marketing (SMM) between B2B, B2C, and mixed business models
11. Effects of Influencers and Sponsorship Prediction of Instagram Posts by them
12. Estimating Tie Strength in Follower Networks to Measure Brand Perceptions

Extraction of dataset

In this task, I used ***Replab*** 2013 dataset for Online Reputation Management, which contains tweet IDs collected by crawling performed during the period from the 1st June 2012 till the 31st Dec 2012. The corpus consists of a collection of tweets referring to a selected set of 61 entities from four domains: automotive, banking, universities and music/artists.

I used twitter API for collection of tweets and some other info about those tweets like the number of retweets, likes, language, timestamp, and if any media links or images were posted with the tweets.

My code used ***Tweepy*** library in python to interact with the twitter API and extract all the info on tweets. Some of the tweets were inaccessible because the account of the tweet was expired or the post was removed from Twitter.

This image below is an example of tweets' info collected from one of the 61 entities explained above.

	tweet_id	text	author	language	likes	retweets	media	timestamp
0	205888692580126720	#radensaleh is not a myth. Learn about his lif...	HelenaAbidin	EN	0	1	no url	2012-05-25 05:11:04
1	207430942028079104	The new BMW 3 Series is awarded 5 stars in the...	VinesBMW	EN	0	2	no url	2012-05-29 11:19:25
2	208204757779759105	@GEAGarratt BMW hand over 200 + electric vehic...	evguide	EN	0	1	no url	2012-05-31 14:34:17
3	208283774251831296	I asked Sauber about more info or images of th...	ScarbsF1	EN	1	3	no url	2012-05-31 19:48:16
4	208342777627549696	Racky must think im driving a BMW or something...	Corey__Deal	EN	0	0	no url	2012-05-31 23:42:43
...	...	...	...	...	...	...	...	...
539	209938458611941376	FREE Black BMW Of Your Choice! Curious! Visit ...	gregmelandow	EN	0	0	no url	2012-06-05 09:23:23
540	208968123179741184	1980 BMW R100RS Cafe Racer http://t.co/golbzeO...	belarba	EN	0	0	no url	2012-06-02 17:07:37
541	210234207379800064	I do agree that money can't buy happiness, But...	eloii19	EN	0	0	no url	2012-06-06 04:58:35
542	209609738441334784	BMW R100/7 by 4H10 Paris http://t.co/dDqbnPHG ...	hrairkhoshafian	EN	0	0	no url	2012-06-04 11:37:10
543	212247976297512960	GM, BMW, Toyota, Audi, Mercedes, Honda, Jaguar...	Pogue	EN	15	139	no url	2012-06-11 18:20:35

544 rows × 8 columns

Fig 1 : - Sample of twitter data collected

Many of the tweets had no media-url. But some of the tweets which had media-url in them were extracted separately and then using **Requests** and **Shutil** library of python the images can be downloaded. The process was automated for all the 61 entities using **Bash script** for automatic creation of folders and filenames of images.

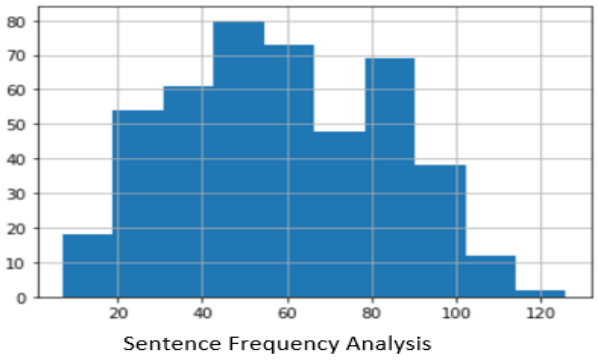
Pre-processing of Tweets

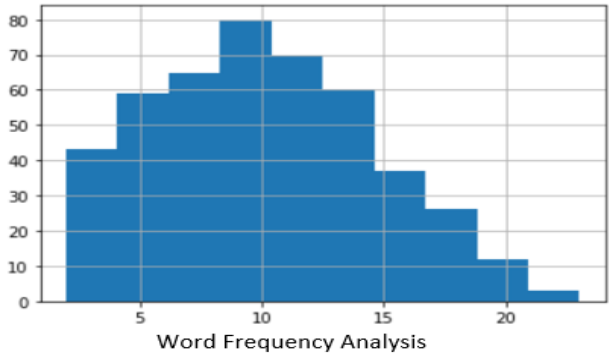
This task was one of the essential steps as the text of tweets in its raw form contains many hashtags (***#radensaleh***), usernames(***@GEAGarratt***), URLs(***http://t.co/dDqbnPHG***), punctuations, and repeating characters in the word(***hellooooo***) with some alphanumeric chars. These things were removed using the ***RegEx*** library while preprocessing as it may produce some wrong results and will consume processing power while doing data analysis and sentiment analysis.

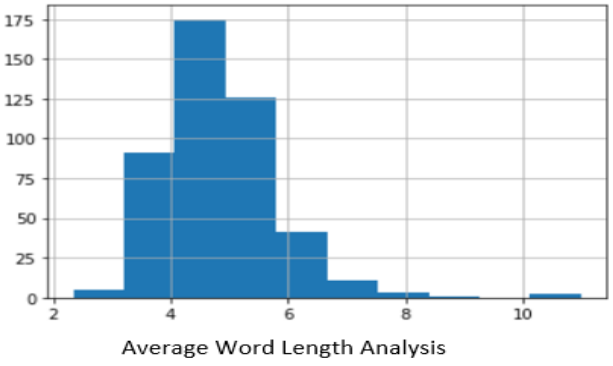
Some of the tweets even contain contracted words (like ***wouldn't***) which were expanded later on using a JSON file containing a list of those words. I removed all the stopwords using ***NLTK*** stopwords library, but it also removed some words which would change the sentiment of the tweet, so I made my own set of stopwords by using some online resources to cover that issue. The dataset also contains some tweets in Spanish which were removed.

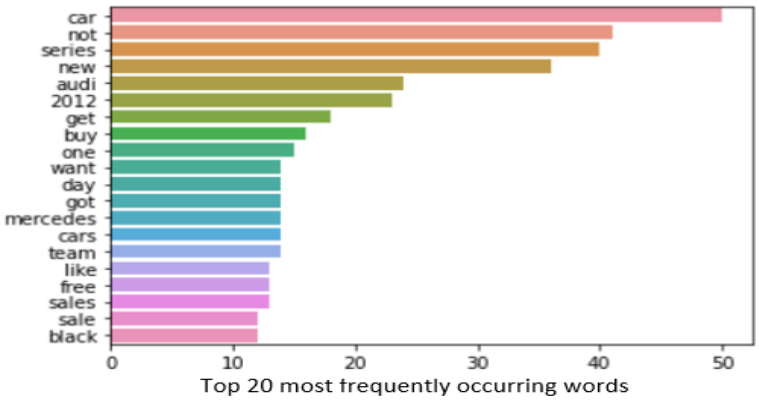
Exploratory Data Analysis of Tweets

1. **Analyzing Text Statistics** - Here we are using text visualizations techniques which are insightful and represent an output of a language model. I used mostly histograms and bar plot charts for representing these analyses. These include: -
    a. Word frequency analysis
    b. Sentence frequency analysis
    c. Average word length analysis



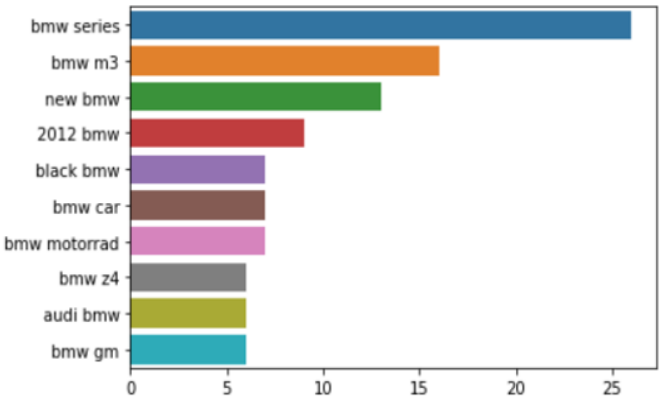




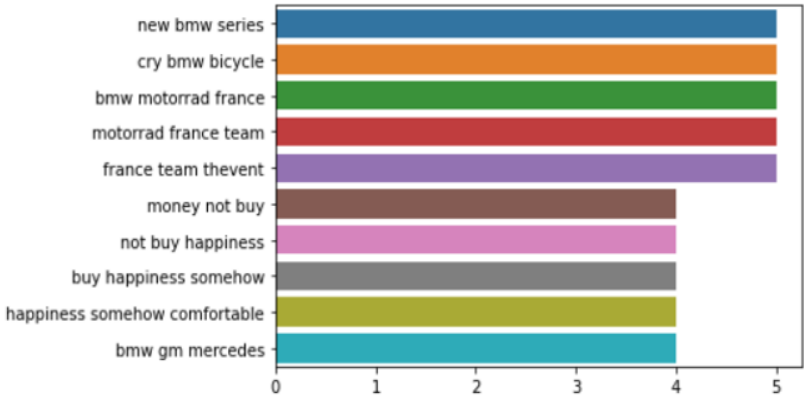


I also used the ***counter*** function from the ***Collections*** library to count and save the frequency of each word in a list of tuples. And, then removed stopwords from them to print a bar plot of top 20 most frequently occurring words which can give us an insight of which is the most discussed topic.

2. **N-gram exploration** - N-grams are a contiguous sequence of n items from a given sample of text. These items can be phonemes, syllables, letters, words or base pairs. For the visualization and the representation of n-grams, I used ***Countervectorizer*** which is used to tokenize, vectorize and represent the corpus in an appropriate form. I did the representation for both the bigrams and trigrams which gave a broader view (for entities like the vehicle, they showed which particular model of the car was on the discussion. Ex: - "BMW m3", "black BMW", "BMW Motorrad France", etc.)



Bigram Plot



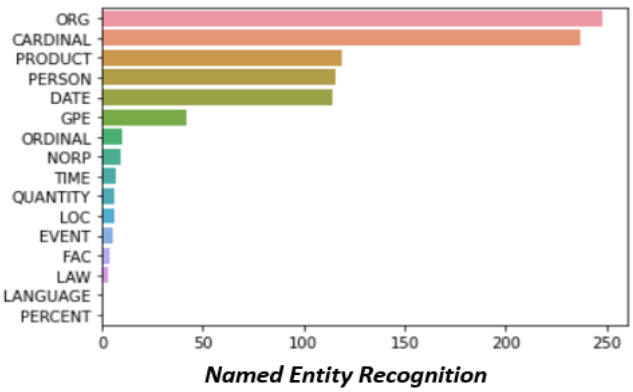
Trigram Plot

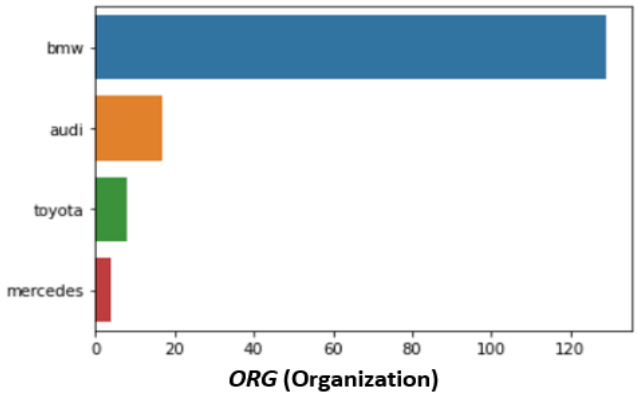
3. **Wordcloud -** Wordcloud is a great way to represent text data. The size and color of each word that appears in the wordcloud indicate it's frequency or importance. Here I created two word clouds, one for the tweets in the text after preprocessing it to remove other repetitive stopwords, which gave an overview of the trending topic in all the tweets and another for the hashtags in the tweet, which would provide us with information like the trending hashtags. But we can observe that wordcloud for hashtags are very scattered and small in font which shows that hashtags are of less importance for this entity.

Tweets' text

Hashtags

4. **Named Entity Recognition –** It is an information extraction method in which entities that are present in the text are classified into predefined entity types like "Person", "Place", "Organization", etc. Using NER, I can get insights like which type of entities are there in the given dataset. I used **spaCy** for NER. We can see that in this dataset, ORG and CARDINAL dominate, followed by the PRODUCT entity. I also visualized the most common tweets per entity for ORG, which can tell us about the brands most popular among those.



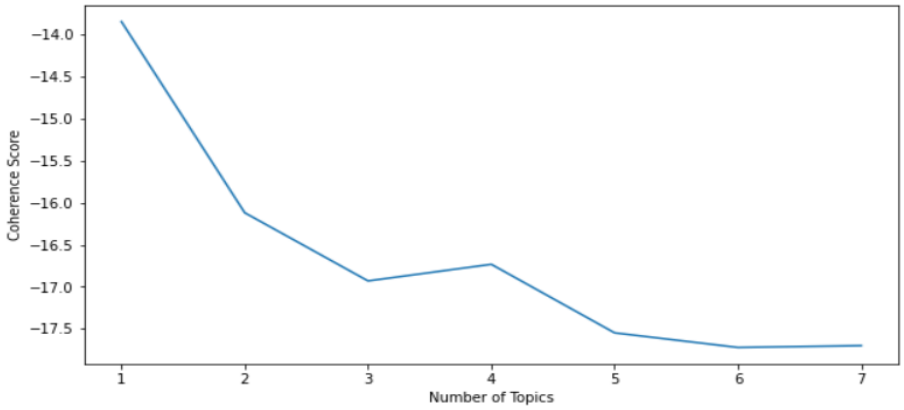


Topic Modelling

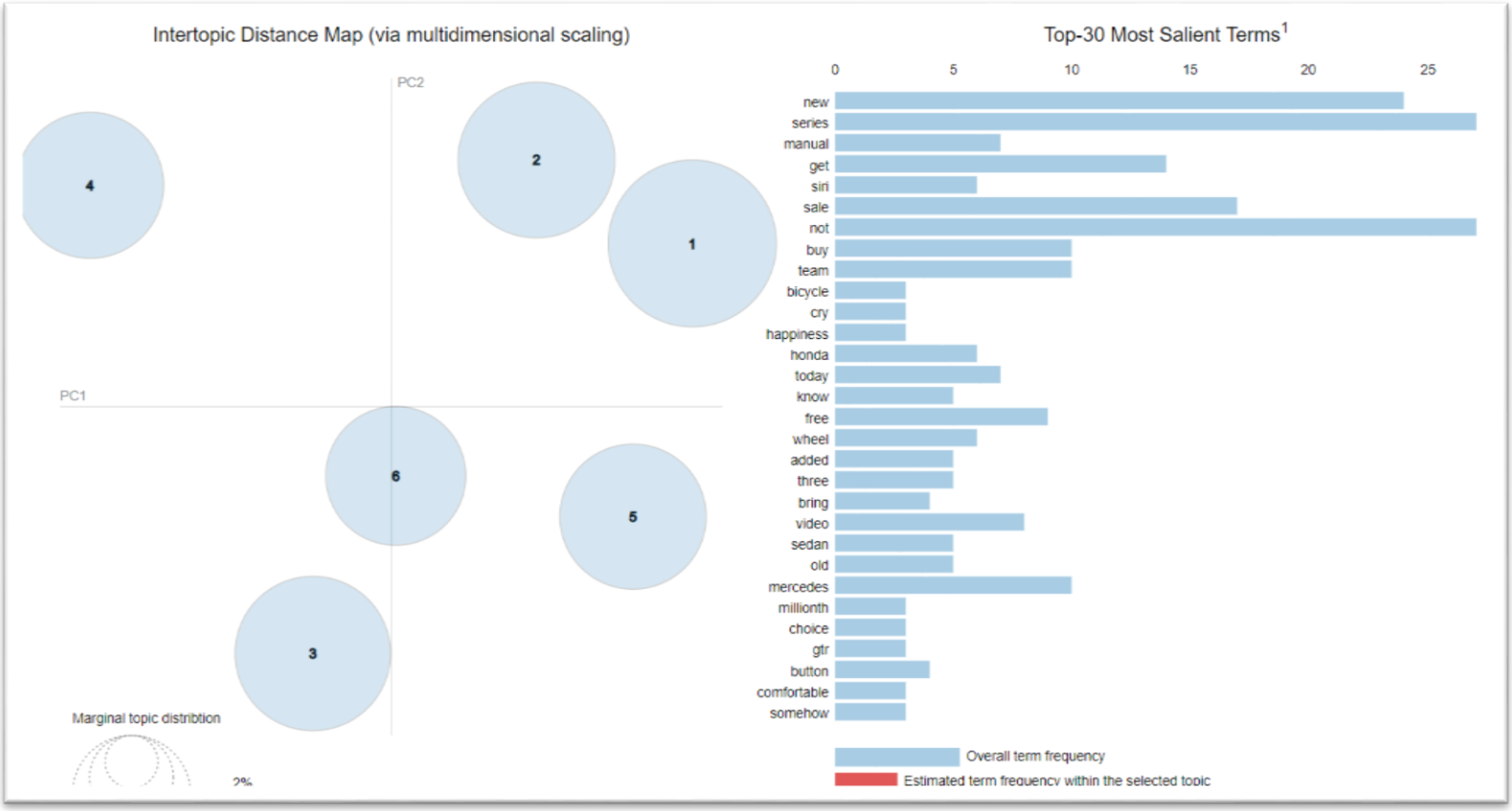
The task was on extracting the main topics from the tweets using unsupervised techniques. I used ***LDA*** (***Latent Dirichlet Analysis***) for topic modelling as it is easy to use and efficient. But before using *LDA*, the tweets need to be pre-processed. I tokenized, removed stopwords, lemmatized and stemmed using the *NLTK* library. And then, after converting them to a bag of words, I created the *LDA* model using Gensim library.

We need to tune the hyper-parameters, i.e. num_topics and $passes$ to get an appropriate number of topics with the most words occurring in that topic to get an insight into the topic. Here num_topics means no. of topics that can be in the document and $passes$ means the no. of loops it will run to determine the core words of the topic

For fine-tuning the hyper-parameters, I manually started with a value of 3 for num_topics and then increasing it manually and checking the values up to 6 or 7 to get a literal understanding of the topics using the core words. The best-suited value for num_topics was 5 which showed meaningful topics covering all the tweets with different core words in them which shows the topics are very much distinct to each other.
But I also calculated the coherence score for the no. of topics ranging from 2 to 8. And the curve of the coherence score showed a minimum at 6 rather than 5, so to cross-check, I used pyLDAvis to visualize both of them. The result shows that for no. of topics equal to 6, the circles were more apart and covered a broader range than 5.



Then, I increased the no. of *passes*, which took more processing power as no. of loops to analyze topics increased to get more accurate core words in the topic.

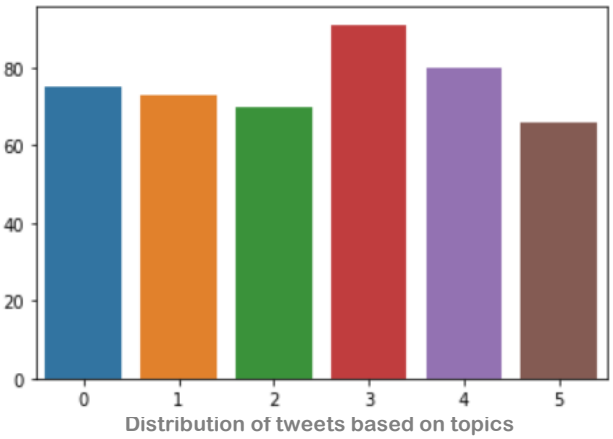


Representation Using pyLDAvis

I used the **pyLDAvis** library to visualize the results of LDA interactively.

- On the left side, **each circle's area represents the importance of the topic** relative to the corpus. As there are four topics, we have four circles.
- The **distance between the center of the circles indicates the similarity** between the topics. Here you can see that the topics are very far apart, and the circles are big, this suggests that the topics are covered over a broad range and are distinct.
- On the right side, the **histogram of each topic shows the top 30 relevant words**.

I have also categorized tweets based on those topics and a bar plot of those categorization would give us an overview of which topic is common in most tweets.

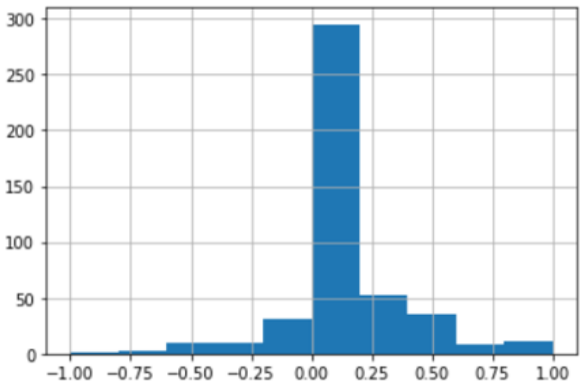


Distribution of tweets based on topics

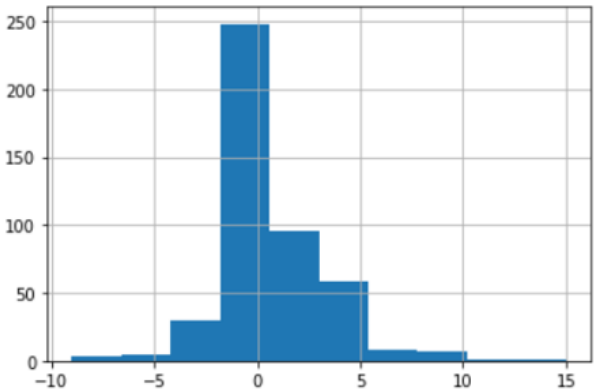
Sentiment Analysis

For sentiment analysis, I used the **Afinn** and the **Textblob** library. Using both the library will give us an intuition, i.e. the difference in the polarity of sentiments for tweets as both the libraries use different algorithms for sentiment analysis.

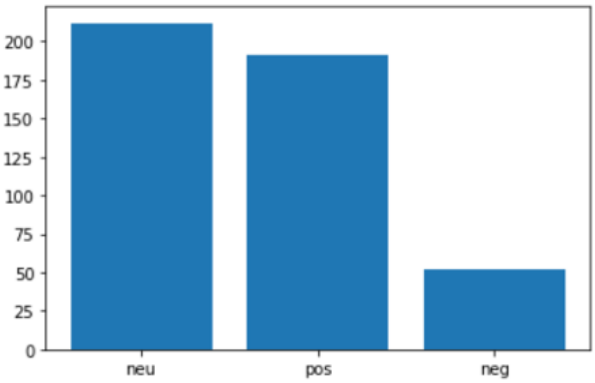
- Textblob library gives us the polarity score and subjectivity score on the tweet. The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0].
- Afinn library gives the polarity score in range [-5, 5] based on the wordlist approach, i.e. it contains the wordlist where each word has a given valence polarity score.



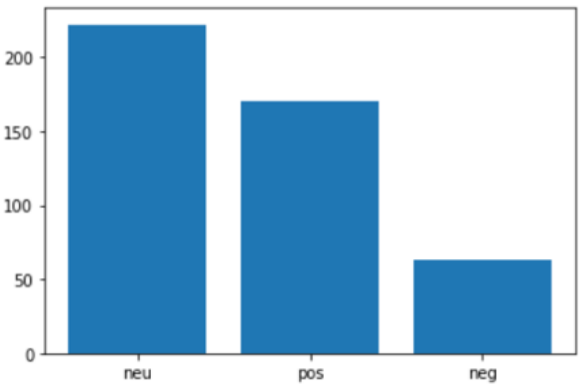
Textblob Polarity score distribution



Afinn Polarity score distribution



Textblob Sentimental Tweet Distribution

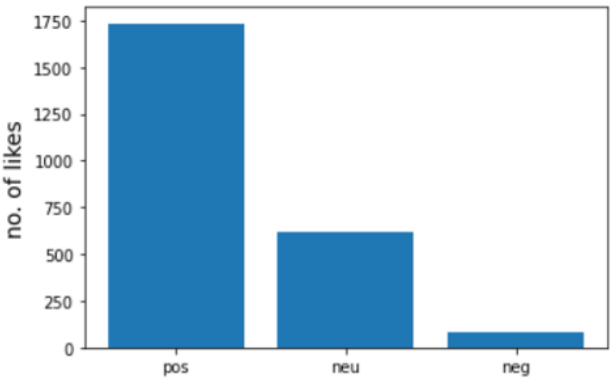


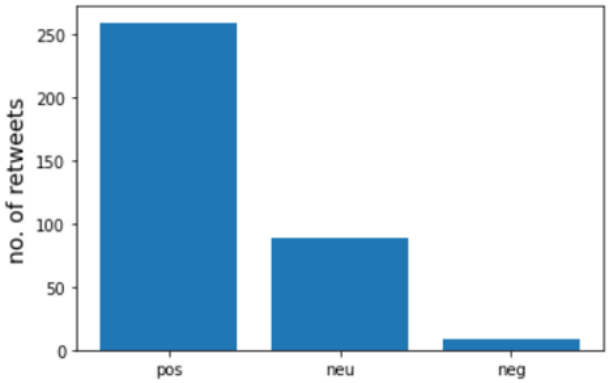
Afinn Sentimental Tweet Distribution

I have also used **Vader Sentiment Analyzer** from the NLTK library. But the results were poor as most of the tweets were categorized as neutral sentiment, and very few tweets were positive or negative, which was not the case.

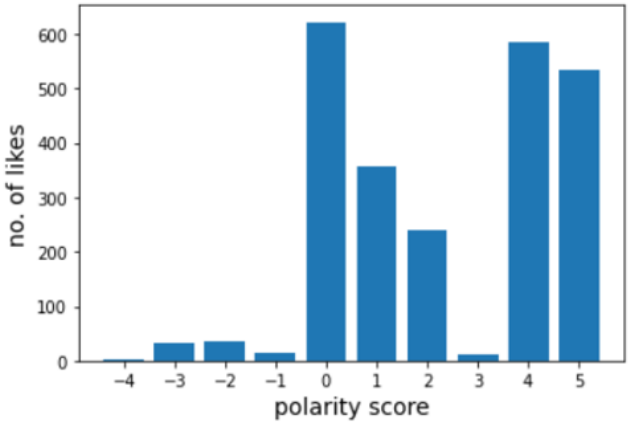
Brand Popularity

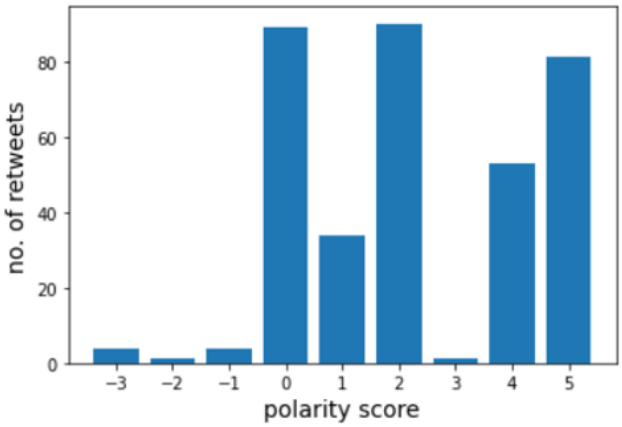
In this task, the main aim was to know people's perceptions about the brand of the products they use. I used the **number of retweets & likes** on a tweet to get an insight into its popularity and the sentiment of the tweet, which would give us an overview i.e. popularity of a particular sentiment about that brand. Using these features, I calculated the no. of likes & retweets for each sentiment (pos, neg, neu).





And, then visualizing it by a bar plot gave us a complete picture of brand popularity. We can also see that which polarity score received the most number of likes and retweets, which will provide us with people's perception in a more detailed way. For the given entity, we can see that no. of likes and retweets on positive sentiment polarity exceeds the negative and neutral counterpart.





- These analyses were done on a single entity (i.e. BMW) , the same can be done on all the 61 entities of Replab dataset.

Brands Similarity

In this part, the main aim was to find the similarity between tweets of different brands of different domains like BMW, Volvo, Royal Bank of Scotland and Barclays. In this example, we can see that BMW and Volvo are from same domains, i.e. Vehicle whereas Royal Bank of Scotland and Barclays is from a different domain, i.e. Finance. So, I took 4 of 61 entities from the dataset to compare the similarity between their tweets.

To compare similarity first, we converted the tweets into vectors using **TfIdf** vectorizer (*I used unigram and bigram to vectorize*) and then used the cosine similarity to compare similarity between these vectors. Here, the values closer to 1 shows that tweets are more similar than the values closer to 0.
I also filtered the tweets by using a threshold value of **cosine similarity** equal to 0.3.

Thus, the results shown that the similarity between tweets of the same domain were more similar than that of between the tweets from different domains.

Tweets 1	Tweets 2	No. of similar tweets(>0.3)	Avg. Cosine Similarity
BMW	Volvo	52	0.474
BMW	Royal Bank of Scotland	58	0.366
BMW	Barclays	58	0.369
Volvo	Royal Bank of Scotland	32	0.365
Volvo	Barclays	37	0.377
Barclays	Royal Bank of Scotland	22	0.386

Here in the above table, we can see that the entities “BMW” and “Volvo” have the most similar tweets as their Avg. Cosine Similarity value is 0.474 whereas the second most similar tweets are for entities “Barclays” and “Royal Bank of Scotland” as their Avg. Cosine Similarity value is 0.386.

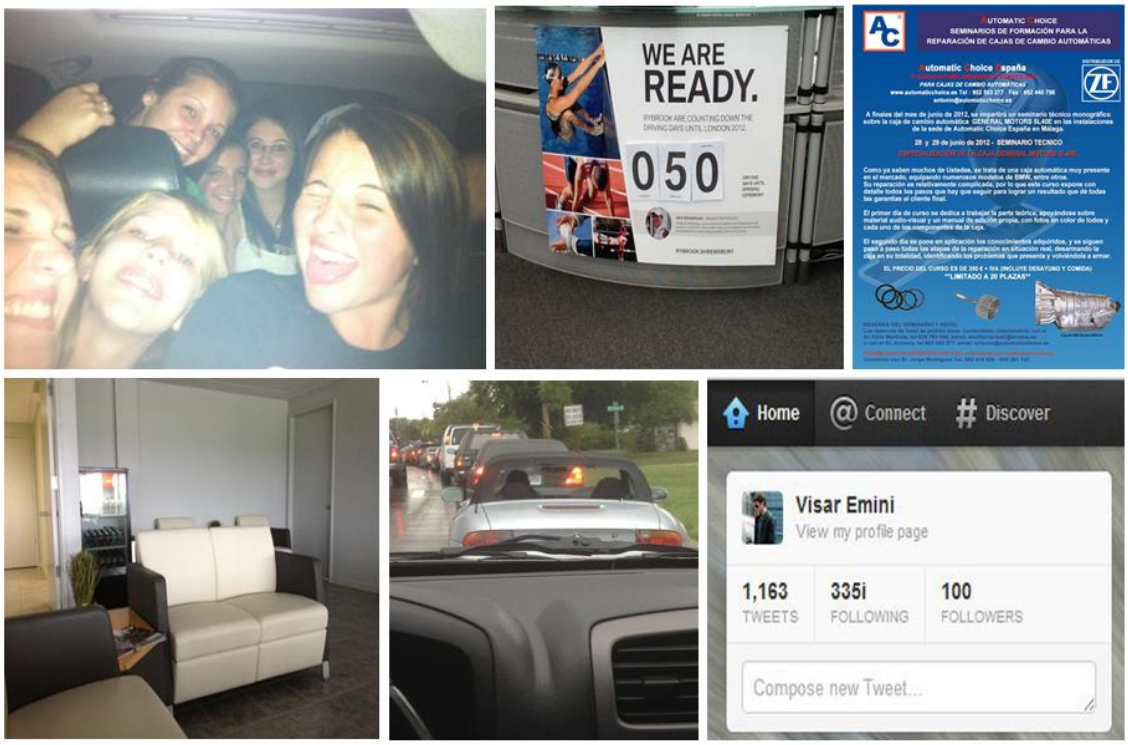
This is self-explanatory as explained above the entities from same domains would be having the more similarity between their tweets. We can also see that the similarity between tweets of different domains are less similar such as “BMW” and “Barclays”, “Volvo” and “Royal Bank of Scotland”, etc.

Given below is a table where we can see that the cosine similarity between tweets of the same entity are calculated with their indices given in the next row.

	tweet_one	tweet_two	cosine_relation	index
0	#for sale oe replacement bmw driver side headl...	#on sale depo 344 1110l asd bmw 5 series drive...	0.492694	15
1	#for sale oe replacement bmw driver side headl...	#forsale tyc 20 6472 01 bmw driver side headli...	0.363815	8
2	#forsale tyc 20 6472 01 bmw driver side headli...	#for sale oe replacement bmw driver side headl...	0.363815	6
3	#forsale tyc 20 6472 01 bmw driver side headli...	#on sale depo 344 1110l asd bmw 5 series drive...	0.300724	15
4	bmw reveals new performance m3 m5 uk market #l...	bmw m3 limited edition uk #1	0.303762	12
...	...	...	...	...
157	dawn mac button siri comes bmw gm land rover a...	dawn mac button siri comes bmw gm land rover a...	0.854956	446
158	dawn mac button siri comes bmw gm land rover a...	bmw gm mercedes land rvoer jaguar audi toyota ...	0.382951	442
159	1980 bmw r100rs cafe racer via	bmw cafe racer	0.637374	243
160	agree money not buy happiness somehow comforta...	agree money not buy happiness somehow comforta...	0.929729	112
161	gm bmw toyota audi mercedes honda jaguar agree...	new feature eves free car manufacturers add si...	0.450275	441

Multimodal Approach

For the multimodal approach, I thought of the images that were extracted. But, on manual inspection of images, I found that 70-80 % of these images mostly contained images of flats, people, car windows, and some other unrelated posters. The same was the case for the images of all the other entities. Thus, the images were irrelevant and they were not contributing to the detection of sentiment analysis and in turn brand popularity of the given brands/entities.



As you can see in the above examples, that these images are very random and won't be contributing to these analyses. Hence, we can conclude that the **70-80 %** of the images used in the tweets are not related to the topic and will be of no use.

Conclusion

This report was based on a comparative analysis of Online Reputation Monitoring. It was done by using the dataset containing tweets on 61 different brands individually. These brands were from various sectors like Automobile, Banking, University and Music Industry. First of all, the pre-processing of the tweets was done then it went through a complete Exploratory Data Analysis of the tweets, which gave us an insight into the dataset. Then, I did sentiment analysis of tweets which then later combined with the number of retweets and likes gave us an overview of the perspective of consumers about the brand. Topic Modelling was done to know the number of topics in the dataset for a particular brand which would give us information about what aspect of the brand is being discussed in the community. I also conclude that the images found in the tweets are very less related to the brand in discussion and thus were not crucial to go for a multimodal approach with them. At last, an analysis of brand similarity was done to find the similarity between tweets of brands of different sectors. This report shows us the number of ways data analysis can be done for Online Reputation Monitoring of brands.

Resources

1. RepLab 2013 Dataset
2. Gensim Tutorial – A Complete Beginners Guide
3. Evaluate Topic Models: Latent Dirichlet Allocation (LDA)
4. Remaking of shortened (SMS/tweet/Post) slangs and word contraction into Sentences NLP
5. Exploratory Data Analysis for Natural Language Processing: A Complete Guide to Python Tools
6. Using LDA Topic Models as a Classification Model Input
7. Estimating Tie Strength in Follower Networks to Measure Brand Perceptions
8. Comparison of different Word Embeddings on Text Similarity — A use case in NLP
9. WASSUP? LOL : Characterizing Out-of-Vocabulary Words in Twitter
10. TexRep: A Text Mining Framework for Online Reputation Monitoring