© https://fti-tn.net/publications FTI Proceedings: 4th international conference on computer applications and information security (iccais’2021) / March 19 / 2021/ Tunisia: https://fti-tn.net/iccais-2021-list-of-papers

Home Office due to COVID-19 - Twitter Data Analysis

1st Victor Ivan Lopez Rodriguez
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, Mexico
a00817161@itesm.mx

2st Hector Gibran Ceballos Cancino
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, Mexico
ceballos@tec.mx

3st Francisco J. Cantu Ortiz
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, Mexico
fcantu@tec.mx

***Abstract*—Due to the covid-pandemic, many employers decided to send home their employees to protect them againstthe fast spread of this virus. There are different factors in which the employees take into account to express their opinion about working from home. Recently, Social Media Networks have evolved as a platform where people express, criticize, and feedback on trending topics of the moment. This work attempts to analyze data from Twitter to study the different opinions and emotions about working from home and focus on finding key information that refutes the analysis. Natural Language Processing techniques, Sentiment Analysis using TextBlob, and a Naive Bayes classification approach and statistical analysis will be performed in this research work.**

***Index Terms*—Sentiment Analysis, Text Mining, Tweets, Python, Na¨ıve Bayes, Classification, WordCloud**

I. INTRODUCTION

Social Networks make it possible for people to share personal information, interests, and opinions. The numberof people using social networks is increasing rapidly. Social Networks are used for faster communication about events or situations in the past, present, and in the future. People always have an opinion (positive or negative) about those events and usually, they express it through social networks.

Today’s current problem: covid-pandemic have to change the way people usually lived their lives. Many companies sent their employees to work in a home office scheme. This has created different opinions about the benefits and disadvantages and people usually express them through a Social Network. These opinions may generate an idea or insight on how Home Office is impacting people’s work life and this can be specially analyzed by companies, technology and furniture business, and educational systems offering courses of focusing at home, stress release among others to take advantage and see people necessities.

According to Social Networks, in this research, we will use Twitter as our main data source of tweets. Tweets are used by users to express or comment about a situation or eventand we will use them in this research for text analysis and natural language processing to generate information about the opinions of Home Office due to the covid-pandemic.

The primary goal of this research is to understand how employees feel about working at home, how it impacts their daily life, their work productivity, and their needs to have a good or equal performance as when they worked at the office.

The research questions stated for this work are the following: How do people feel about home office on Twitter? Tweets usually express opinions but can someof them be considered facts? Can Sentiment Analysis be implemented using a Na¨ıve Bayes approach? What are the most used hashtags? Can term frequency give meaning to tweets?

In Section 2 we present the methodology of this research work that includes Background, Theoretical Framework, Data Extraction, Data Preparation and Cleaning, Developments, and Results. In Section 3 we carry out an analysis of methods and results. Finally, conclusions are stated in Section 4.

II. METHODOLOGY

A. Background / Theoretical Framework

Twitter is a social media network that has been increasing its numbers of active users. With the impact of technology, Twitter has become a fast way of communication in terms of news, events, situations in which people share their opiniononit. To share their opinion, Twitter has a concept named Tweet in which a user can express their-selves witha 140 character message. A Tweet can have a symbol HASH, colloquially named hashtag, in which users use the concatenation of this symbol and a relevant word to categorizethe tweet to a proper event, situation, person, among others.

In order to work with data, it is important to apply Natural Language Processing (NLP) techniques to have a clean data set and avoid any type of noise. Manning et al [1] describe some common NLP techniques:

- Tokenization: Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation.

- Stop Words Removal: Common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely.

- Stemming and Lemmatization: The goal is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

Twitter data can be analyzed in several ways to understand the user's opinions. Text Analysis, Text Categorization, and Time Analysis can help to reach the goal of this research work. Text Sentiment analysis is an automatic process for determining whether a text segment contains objective or opinionated content, and it can furthermore determine the text's sentiment polarity [2]. The objective of applying this method on Twitter is to obtain the polarity if a tweet ispositive or negative and the subjectivity if a tweet is a fact or an opinion according to the tweet. Text categorization is the process of automatically assigning one or more predefined categories to text documents [3]. The goal of applying categorization on tweets is to obtain information about age, region, domain and, business sector impact where the tweetis impacting.

In this research work, we are also testing Na¨ıve Bayes classifier on the sentiments approach. Na¨ıve Bayes classifier is a probabilistic classifier based on Bayes' theorem, which assumes that each feature makes an independent and equal contribution to the target class. It assumes that each feature is independent and does not interact with each other, such that each feature independently and equally contributes to the probability of a sample belonging to a specific class. NB classifier is simple to implement and computationally fast and performs well on large datasets having high dimensionality [4].

Social Network analysis has been raising its popularity in recent times because it is a way to obtain feedback about situations, events, products, among others. As technology evolves, users in social networks are increasing their use and companies are using this key variable to advertise or have a better communication channel with their clients.

Due to the increased number of users, Twitter has thesocial media network the people prefer for obtaining, sharing, and spread important information of big relevance such as covid-pandemic [5]. During all the period in which we have been in this pandemic, a large quantity of tweets has been generated and used for different study cases as a source of information in which topics as content analysis, sentiment analysis, forecasting, detection of outbreaks among other studies [6].

B. *Data Extraction*

To obtain the information from Twitter. We will use Twitter API to facilitate the procedure of gathering information. Twitter API (Application Program Interface) let you readand write Twitter data and access to high volume tweets ina particular subject.

The first step of the data extraction is to create a Twitter developer account. For applying for an account, it is needed to specify the reason for using the Twitter API and explaining in detail the analysis proposed. After completing the application, it is reviewed. Once the application is accepted, we will use the REST API to access, search, filter, and download the tweet data we need for our project.

After having an approved account and log in to the Twitter API Dashboard, it will automatically guide you to createan App and inside it, create a Project. When the project is created, it will generate two User Access Keys and Tokens. It is important to save them because they would be needed in the next steps.

For the data extraction, python is going to be used with the tweepy library. Authentication code is needed with the previous keys and tokens mentioned. Afterward, a dataframe using pandas library in python will be created and the api.search method from the tweepy library makesthe search with the following parameters: tweet contains the word 'homeoffice', tweet language is in English, discardingretweets to avoid the repetition of tweets, and a maximum amount of 2000 tweets. Twitter API with a developer subscription only allows retrieving tweets from a week ago when the query is performed. There are different alternatives such as private platforms that allow you to retrieve historical data. In this research work, we are going to analyze tweets with the developer subscription.

C. *Data Preparation and Cleaning*

After the data extraction, we need to prepare andunderstand the information to apply the methods on it. With the source data, the process continues in understanding how the data is right now and the data is needed for applying the analysis methods.

The tweet comes with a lot of information and it is necessary to clean it. We will take out all the elements that may corrupt the main focus of the expression in thetweet.This process will be made automatically after the data extraction gets the information from Twitter API.

An element in a tweet, such as URLs (links) and images do not add value and will be noisy when we start making statistics and data analysis. The first step is to lower case the text in the tweets and split compound words. Lower casing the text helps us in avoiding possible differences in further methods of categorization and polarity.

The method used for punctuation removal in our python program is using the regular expression library. Some defined regular expressions functions detect if there is an URL in the tweet and if it is true, it replaces it for an empty space " ". The same regular expression also analyzes symbols as hashtags and replace them with an empty space. Images are not considered in our search methods for tweets, therefore there is no need to apply another method.

After removing punctuation and special characters, such as Twitter Acronyms and URLs that do not add value tothe tweet, the following step is to tokenize the tweet usinga regular expression in python. Another method applied for cleaning up the data is to take out the stop words. Stopwords are do not add meaningful information to the text.For this step, the python nltk library will be used.

As tweets express a personal opinion, there are several ways to express it. Emojis represent support for the tweet. Depending on the type of emoji used, we can give us an idea if the expression is positive or negative. Nevertheless, emojis will not count for our analysis because they can also tend to be express sarcasm or metaphors that do not enter the scope of this research work.

All the techniques are applied to each tweet obtained from the data extraction and will return a clean data set ready to be worked with.

D. Development and Results

After completing the cleaning phase, the research work continues with the implementation of several methods to understand what the information says. The first method is Sentiment Analysis. The library used is TextBlob, it enables your code to perform operations such as Sentiment Analysis, Tokenization, Lemmatization, Noun phrase extraction among others on textual data. Two functions are created using the TextBlob property sentiment which returns a tuple (polarity, subjectivity). The first function is used to get the polarityof the tweet. Polarity means the emotion expressed in the tweet and it can be negative or positivein a range from [- 1.0, 1.0]. The second function is usedto get the subjectivityof the tweet. Subjectivity means ifa tweet expresses a factor or an opinion in a range from[-1.0, 1.0]. Both functions are used in each tweet in an iteration of all the tweets(already cleaned up) in the data set.

The next step is to import the seaborn and matplotlib libraries to have an interface for plotting the processed information.

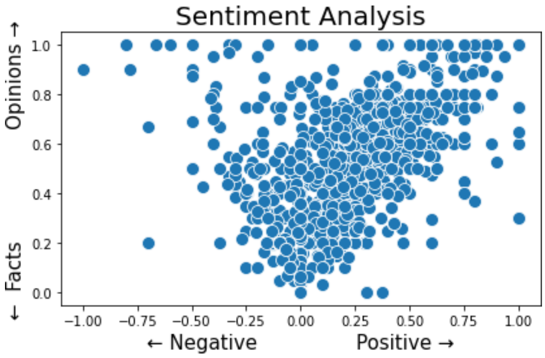


Fig. 1. Sentiment Analysis of home office tweets

Afterwards, it is important to plot each attribute of Sentiment Analysis in its own as a Histogram in order to analyze deeper each concept.

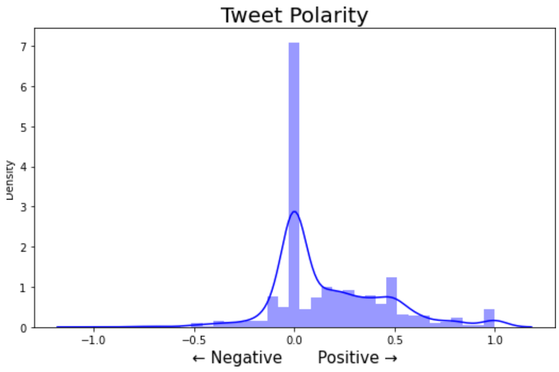


Fig. 2. Polarity of home office tweets.

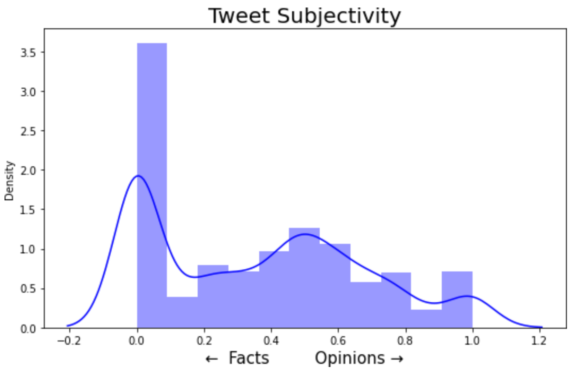


Fig. 3. Subjectivity of home office tweets.

In this approach, we are not considering the number of retweets and likes a Tweet has because we are deleting duplicates. This value may impact polarity in terms of frequency in a period of time and in this research work we are evaluating how polarity is by tweet.

The next analysis is a Word Cloud that evaluates the term frequency in all the data set to give us an idea of common words used in tweets related to home office. For this process, the WordCloud library in python is imported and used. A generation of a raw string which concatenates each word of all tweets (already cleaned) in the original dataframe and the string was passed to the function to process it and plot the following image.

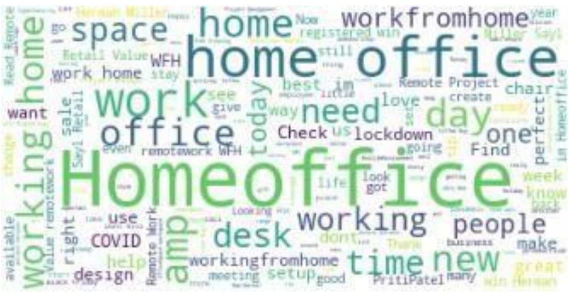


Fig. 4. WordCloud of home office tweets.

Another analysis performed using term frequency is the occurrences of hashtags in home office tweets. Hashtags are considered a valuable property of a tweet because users use them to focus on and emphasize the content of the tweet. A new dataframe was created to modify it in terms of frequency of hashtags or bigrams.

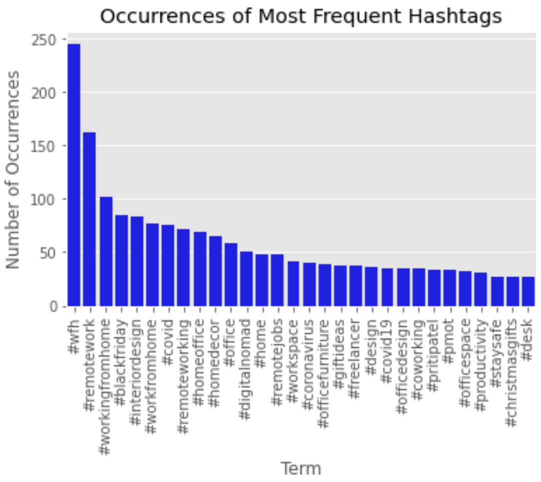


Fig. 5. Frequency of Hashtags in home office tweets.

The next analysis is about Time Series, a new dataframe is created with the same data of the original dataframe tweets of home office. This new dataframe is indexed bythe column "created at". This column has information about whenthe tweet was created. With this index, we are creatinga graphin which the information is analyzed in terms ofdays and the polarity of a tweet. This information is valuable because we can see how polarity value is going in timeduring all the lockdown and important months of the year.

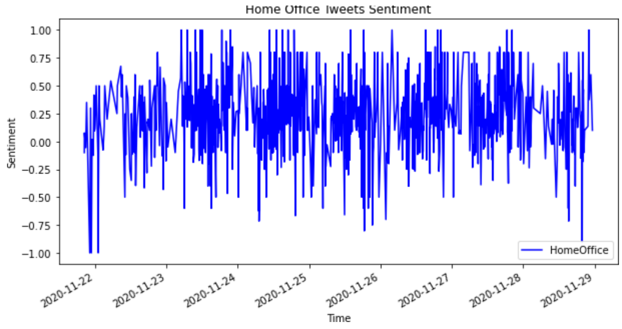
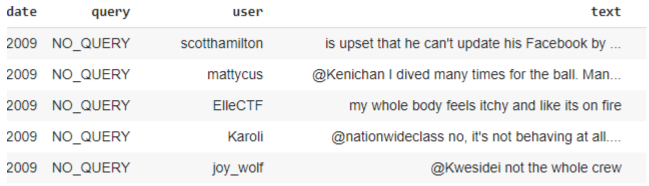


Fig. 6. Time Series of home office tweets in a week.

As we mentioned before, the Twitter API only let developers download tweets from a week before, and for this kind of analysis, it is better to have a historical dataset of tweets related to home office since the beginning of the pandemic until now to perform other kinds of analysis such as trends and seasonality.

.

The last method to apply in the analysis of home office tweets is an approach of sentiment classification using Na¨ıve Bayes. A dataset retrieved from Kaggle used in a similar approach developed by Go et Al [7] which contains 1.6 million tweets extracted from the API in which they have a sentiment classification of 0 meaning negative, 2 meaning neutral, and 4 meaning positive. It has fields as sentiment, tweet id, date, user, and text. The tweets do not have a specific query and it will be managed as historical data for training.



date	query	user	text
2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....
2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew

Fig. 7. Training Set of Twitter Data.

The first step was to upload the mentioned dataset to our jupyter notebook and it will be read as a dataframe to be used as our training set. As we are plotting sentiment using TextBlob and the value range that it manages is between -1 and 1, the following step is to standardize this value in the training set as we manage the value in TextBlob. The library natural language toolkit (nltk) to use the Na¨ıve Bayes Classifier to extract the features, process the training set and train the model. After the model was trained, an iteration overthe test set (the dataset with the home office tweets) was run to calculate the new polarity based on the Na¨ıve Bayes model.

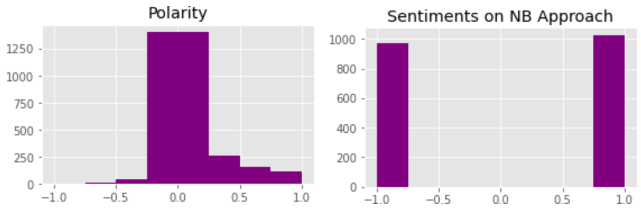


Fig. 8. Comparison on Sentiment Analysis (TextBlox and NB Model)

The classification of tweets shown in Figure 8 using the Na¨ıve Bayes model did not classify neutral tweets as it considered, by the extracted features of the training set, that all the tweets were positive or negative. It may impact that tweets are not related to a specific topic in the trainingset and emotions may be express in other ways. The main scope of this approach was sentiment classification and it succeeded but results can be improved using historical dataof homeoffice tweets. TextBlob had better results in terms ofaccuracy as it returns decimals to see its classification.

III. DISCUSSION

Sentiment Analysis implementation in extracted tweets including home office as its main search concept gave interesting results that can be analyzed with the dispersion in Figure 1 and histogram graphics in Figures 2 and 3. In the result of the dispersion graph it is stated that most of the tweet tend to have more neutral and positive than negative polarity and that almost all the tweets dispersed tend to be an opinion than a fact. We can deduce from the graph the most of the people express positive emotion by working at home.

The histogram of tweet polarity shows that there is a high density in neutral tweets. This result can tell us that we should improve the way we are cleaning up the information, perhaps at the cleaning stage, we are taking out keywordsthat express the opinion of that tweet. In this histogram,we cansee that there is more density in the positive range of polarity than negative tweets. The histogram of tweet subjectivity expresses the results if a tweet content is an opinion or a fact. It tends to have a big density of neutral tweets. As same as the polarity histogram, we should also improve the cleaning stage. As tweets tend to be an opinion of the user concerning a topic or situation, it is correct that this Sentiment Analysis has more density in the opinion range.

In the wordcloud method, we can see the words with more frequency in the test set. Words such as best, setup, time, great, remote, need, space, perfect, good, and covid can define an important idea the reason for the tweet sentiment These words can let companies know what is affection workers opinions and work in a strategy to attack this problem. The setup in a home office workplace plays an important role in productivity and companies could provide headphones, laptops, desks, among other things to fulfill the necessity. Theoccurrence of Hashtags can help to identify other missing dataas they are acronyms and alternative ways to name home office as work from home (wfh). Other hashtags used are about covid19, remote work, coworking, office design, productivity, and workspace.

The analysis of tweets over time can let companies know how people is feeling during all the pandemic. In Figure 6 we can see a complete week evaluated from Sundayto Sunday. Tweets present more negative sentiment on Sunday and Monday and can be related to the energy of the beginning of the week.

The Na¨ıve Bayes sentiment classification approach was correctly applied, nevertheless, the results were not as good as expected. It did not classify any neutral tweet and TextBlob had a big quantity of neutral tweets. A reason that can explain these results is that historical data is not relatedto home office tweets, language can be more technical andthe waythe library nltk uses the training for the Na¨ıve Bayesclassifier.

IV. CONCLUSION

After analyzing all the results of the methods applied to the source information. We can conclude that people, in general, feel happy and comfortable working at home during this difficult time due to the covid pandemic. Thanks to the evolution of technology and the internet, most of the jobs can be done at home.

Word Frequency and Hashtags occurrences results can state that people usually express opinions about the time, workspace, and setups in which they work every day. This can be a good area to analyze as an opportunity for companiesto make strategies that may impact productivity.

Na¨ıve Bayes approach was a good experiment by changing the scope of classifying sentiment instead of tweet concepts. This research work can continue in the future with the analysis, comparison, and correlation of different search terms such as home office and work from home. An analysis of English and Spanish tweets and gather historical data of this topic to perform time series analysis such as trends and seasonality.

Nowadays social media can be considered a big tool for data analysis as there are big opportunity areas to exploit.It can analyze opinions, interests, product impact, and other critical data that can be used for marketing, feedback, and new strategies. We expect that this research can be used by companies or institutions in any field to benefit society. The generated knowledge is intended to be shared and used for improvements in future work.

.

ACKNOWLEDGMENT

We gratefully want to acknowledge Dr. Hector Ceballos-Cancino and Dr. Francisco Cantu-Ortiz for their feedback and revision and thank Tecnologico de Monterrey and CONACyT for the financial support.

REFERENCES

[1] Schütze, Hinrich, Christopher D. Manning, and Prabhakar Raghavan. Introduction to information retrieval. Vol. 39. Cambridge: Cambridge University Press, 2008.

[2] Jianqiang, Zhao, Gui Xiaolin, and Zhang Xuejun. "Deep convolution neural networks for twitter sentiment analysis." IEEE Access 6 (2018): 23253-23260.

[3] Billsus, Daniel, and Michael J. Pazzani. "User modeling for adaptive news access." User modeling and user-adapted interaction 10.2-3 (2000): 147-180.

[4] Sciencedirect.com. 2020. Naive Bayes Classifier - An Overview — Sciencedirect Topics. [online] Available at: ¡https://www.sciencedirect.com/topics/engineering/naive-bayes-classifier¿ [Accessed 29 November 2020].

[5] Rosenberg, Hans, Shahbaz Syed, and Salim Rezaie. "The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic." Canadian Journal of Emergency Medicine (2020): 1-4.

[6] Gencoglu, Oguzhan, and Mathias Gruber. "Causal Modeling of Twitter Activity During COVID-19." arXiv preprint arXiv:2005.07952 (2020).

[7] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N project report, Stanford 1.12 (2009): 2009.