uc3m | Universidad Carlos III de Madrid

Master Degree in Computational Social Science
2022-2023

Master Thesis

Cross Border Twitter Analysis: Identifying International Customer Interest

Yassin Mohamed Kamel Ahmed Abdelhady

Supervisor: Sebastian Daza

Madrid, 2023

AVOID PLAGIARISM

The University uses the Turnitin Feedback Studio program within the Aula Global for the delivery of student work. This program compares the originality of the work delivered by each student with millions of electronic resources and detects those parts of the text that are copied and pasted. Plagiarizing in a TFM is considered a **Serious Misconduct**, and may result in permanent expulsion from the University.

[Include this code in case you want your Master Thesis published in Open Access University Repository]
This work is licensed under Creative Commons **Attribution – Non Commercial – Non Derivatives**

Abstract

The food sector is about $2,323.29 billion, however the vegan food market only accounts for 1% of that total. Social media, especially online communities, is an effective tool for connecting businesses with their target market. Utilizing social media can help startups grow into new markets and make educated judgments. This study uses topic modeling on Twitter data to find potential consumers from a Spanish company that operates in the teritory of Spain to Germany. The research design entails gathering tweets from the followers of the Spanish company and subjecting them to topic modeling methods of analysis. For contrast, the tweets from Berlin's SBahnBerlin are also examined. The data is processed, including lemmatization, translation, cleaning, and encryption of sensitive material. The parallels between the followers of the Spanish company and the followers of SBahnBerlin are examined through topic modeling using BERTopic. The drawbacks include dependency on processing power, noise in Twitter data, and a lack of contextual awareness. The study's findings, which include 199 tweets from SBahnBerlin followers and 123 Berlin users interested in veganism, point to low possible commercial opportunities.

Public repository for the source code:

https://github.com/yassinabdelhadyy/cross-border-twitter-analysis-TFM

Social Network – Topic Modelling - Twitter - startups - Vegan

Contents

INTRODUCTION

The food industry is an enormous market, with a value of 2,323.29 billion USD in 2021 (Food Service Market Size, 2023). In contrast, the Vegan food market size was valued at 26.16 billion USD in the same year, representing only 1% of the industry. These figures clearly indicate the substantial potential growth in the vegan food sector.

In today's digital era, social media has emerged as a powerful tool for businesses to connect with their target audience, build brand awareness, and ultimately penetrate the market (Jagongo & Kinyua, 2013). Communities have been an integral part of human existence since time immemorial. With the advent of the internet, online communities have emerged as powerful platforms that significantly enhance the connectivity between individuals and businesses in the virtual sphere (Harris & Rae, 2009).

Startups play a crucial role in driving economic growth across Europe. It is worth noting that only a small fraction, around 1% of the companies, have more than 50 employees (Morrison et al., 2003). When considering expansion into new geographic locations, making the right decisions becomes even more crucial. Hasty decision-making can have severe consequences, potentially leading to the company's downfall (Chung et al., 2007). Fortunately, startups can significantly improve their decision-making process by adopting straight forward techniques. Rather than investing in costly market research or relying on luck, startups can harness the tremendous power of social media, primarily if they have cultivated a dedicated community on these platforms.

This paper aims to identify potential customers for a company looking to expand its reach to new regions or cities. For this purpose, I selected a Spanish plant-based company to determine potential customers abroad, and help to make informed decisions. The target city chosen for this study is Berlin, considering that Germany recorded 1.91 billion euros in plant-based food sales in 2022 (state of the industry report Germany 2022), surpassing Spain's 441.7 million euros in the same category (state of the industry report Spain 2022).

Analyzing the followers of the Spanish company Twitter account can serve as a reference point for developing a topic modeling algorithm. This algorithm will then be applied to Twitter accounts in Berlin to determine whether the interests of people in Berlin align with those following in Spain.

In the upcoming sections, the paper Will highlight a detailed overview of the research design for conducting a comprehensive topic modeling analysis. This analysis will be performed using Python and BERTopic, a powerful library for topic modeling. To begin, the methodology used to gather, preprocess, and analyze the data. This involves collecting relevant data from reliable sources and applying various preprocessing techniques such as text cleaning, tokenization, and removing stop words. Next, Modern language representation models are used by BERTopic to create relevant topics from the provided text input. This method enables a more subtle comprehension of the underlying themes and ideas contained in the data. The utilization of Natural Language Processing techniques which BERTopic uses, specifically the class-based term frequency–inverse document frequency (c-TF-IDF). This method enables the identification of key terms and concepts within the text, facilitating a deeper exploration of the dataset and enhancing the accuracy of the topic modeling analysis. Finally, the paper will conclude by highlighting the main conclusions drawn from the analysis. These conclusions will provide insights into the discovered topics and their relevance to the problem at hand. Additionally, any significant findings or patterns that emerge from the analysis will be discussed, contributing to a comprehensive understanding of the dataset and potentially offering practical solutions to the problem being addressed.

RESEARCH DESIGN

Firstly, the tweets of users from the Spanish company will be collected using twitter API to obtian the relevent data. Those tweets will then undergo a process of encryption for the user data, text cleaning, and analysis. Using topic modeling techniques to classify the various topics discussed. Similarly, the tweets of SBahnBerlin from Berlin, will be retrieved and subjected to the same analysis. By extracting the topics from the tweets of the Spanish company, the focus will be placed on assessing these topics and selecting the desired topic, such as "vegan." This selection will enable the identification of potential new customers in Berlin who share same interest in veganism.

The training population for this study comprises 3,998 Twitter users who actively engaging with the Spanish company account on Twitter. To gather relevant data for analysis, a total of 325,098 tweets were collected from these followers within a specific timeframe, ranging from March 27, 2023, to April 26, 2023. This particular timeframe was chosen due to the company launched a crowdfunding campaign on April 21, 2023. By including tweets from this period, the study aims to capture both a normal level of user activity and any potential increase in engagement near the time of the crowdfunding event.

Moreover, SBahnBerlin has a substantial follower base of approximately 240,000 users. However, for computational efficiency and to focus on more active users, certain criteria have been implemented to narrow down the selection. A calculated variable was devised to assess user activity per month by considering factors such as account creation date and the number of tweets posted per month. A threshold of 20 tweets per month was set to identify consistently engaged users. To further refine the sample, accounts with protected settings were excluded, ensuring that only public data was collected. Additionally, a prerequisite was established for the account to be active for at least six months. These measures helped to streamline the study's focus on users who regularly participate and engage on the platform. Consequently, the number of SBahnBerlin followers considered for data collection has been reduced to 4,457 followers and 420,762 tweets gathered from the SBahnBerlin account within the same specified time frame.

By comparing the two sets of data from the Spainsh company followers tweets and SBahnBerlin followers tweets in Berlin, the study aims to analyze any notable similarities in topics. This comparative analysis will provide valuable insights into whether berlin is a good market for planet-based food companies .

METHODOLOGY

To acheive the objective of the study tweets from the twitter accounts was retrived from the target company and the target market. Python Will be used as the main programing language since all the packages needed to complete the study is well supported by Python and are up to date by the community. Harnessing the power of Python with the open-source community of hugging face to use the translation pre-trained models from transformers and the bertopic for topic modeling with some data pre-processing before hand like encrypting the senstive information such as the user id and the tweet id to cross out any path for user information retrival.

- Data Collection

A collection of tweets was gathered using Twitter's API from the Spanish company followers during two distinct periods, resulting in a total of 325,098 tweets. Unfortunately, the API do not provide the retrieval of tweets using location. To overcome this limitation, an alternative approach was adopted. A search was conducted for a Twitter account that residents of the target area, Berlin, would likely to follow. SBahnBerlin is the official account for Berlin's transportation system it was selected as the target account. Given its regular updates on the transportation system in Berlin, it is expected to have a significant number of followers who reside in the city which resulted in 460,932 tweets. Twitter API retrieve a lot of information besides the tweets, since we are only interested in the tweets itself only the id of the tweet and the text and the user id will be kept for the analysis.

- Pre-Processing

Given that this paper focuses on analyzing tweets, certain precautions must be implemented to protect the anonymity of the individuals posting them. To achieve this, an encryption has been conducted. As every tweet possesses a distinct ID, the original ID of the tweet would be encrypted with a combination of a static key that will not be published and the original ID of the tweet, this key is a randomly generated once consisting of 44 characters with all the posible characters can be added for example “Ab!$bA12_” and this charcater string is turned into bytes and added to the original id and they both get encrypted and it is URL-safe base64-encoded. This encryption cannot be altered or read without the key. This same method of encryption Will also be applied to the user id to mask the identity of the user, this will elimnate any personal data that can be traced back to the user.

The utilization of Transformers facilitated the translation of tweets from Spanish to English and German to English. This machine learning model encompasses open-source pretrained models for a wide range of languages, enabling the translation of text from the target language with the aid of frameworks such as PyTorch, TensorFlow, and JAX. It is necessary to align the tweets to a consistent language. Since the Spanish company followers predominantly speak Spanish and SBahnBerlin's followers primarily speak German, both sets of data were aligned to English text to ensure more accurate results.

Given that we are working with tweets, a thorough cleaning process is essential before conducting the analysis. Emojis, retweets, mentions, numbers, links, empty spaces, and line breaks were removed to obtain plain text that can be analyzed effectively.

- Processing

Given that the text is now in English to further refine the analysis the English stop words should be removed like (I, Me, My , Myself,.. etc) they consist of 179 words from the nltk package in python, after analyzing the text custom stop words has been added to the list like (the, and) by removing the stop words the text remaining become more valuable since those words are just connecters and the sentence would be still understandable without them.

For further refining the dataset tweets with less than 3 words were removed so tweets with no text now will be removed and tweets with less than 3 words, which has decreased the tweets from 325 to 300 for the Spanish company and from 460,932 to 420,961 for SBahnberlin.

Then after the stop word removal and word count a technique called lemmatizing was performed on the text to find the normalized form of the word, it is one of the key components of Natural Language Processing and text mining, this process is almost like stemming but when lemmatizing the word the prefix and suffix of the word for instance working, works, worked would change to get the normalized form work standing for the infinitive: work; in this case, both the normalized word form and the word stem are equal. Sometimes the normalized form may be different than the stem of the word. For

example, the words computes, computing , computed would be stemmed to comput, but their normalized form is the infinitive of the verb: compute (Plisson, Lavrac, & Mladenic, 2004)

To analyze the data we utilized the power of the open-source community and we used Bertopic as the main topic modeling technique that embrace transformers and class based term frequency inverse topic frequency “c-TF-IDF” to create dense clustering of the topics which is better since twitter is a micro-blogging social media platform with a maximum of 140 characters per tweet.

Corpus

A corpus will be all the text collected for each twitter account so in this case we will have 2 corpuses and each corpus has text with each text has a unique identifier for the tweet those unique identifiers will be treated as the document number. We will take 2 tweets for an example to explain the following part.

This whole 2 documents “tweets” will represent a corpus with a D

Table 3.1

Corpus D	
d1	understand, people usually agree really, life same.
d2	remember people say see someone gym "you much muscle," "you better before," play hair body apply overweight. thank you.

TF-IDF

TF is Term frequency of any term in a given document

Table 3.2

Corpus D		TF word “people”
d1	understand, people usually agree really, life same.	TF(“people”,d1)=1/7 = 0.142
d2	Remember people say see someone gym "you much muscle," "you better before," play hair body apply overweight. thank you.	TF(“people”,d2)=1/19 = 0.052

IDF is the constant per corpus, and accounts for the ration of documents that include that specific term(Ramos J,2003).

Inverse document frequency

Table 3.3

Corpus D		IDF
d1	understand, people usually agree really, life same.	IDF(“people”,D) = log(2/2) = 0
d2	Remember people say see someone gym "you much muscle," "you better before," play hair body apply overweight. thank you.	IDF(“people”,D) = log(2/2) = 0

TF-IDF

TF-IDF = (word,dn,D)

So for the previous example the word people would be calculated as follows

(“people”,d1,D) = 0.142 * 0 = 0

(“people”,d2,D) = 0.052 * 0 = 0

The word people is equally relevant for both documents in d1 and d2

c-TF-IDF

The goal of the class-based TF-IDF is to supply all documents within a single class with the same class vector. To do so, we have to start looking at the TF-IDF from a class-based point of view instead of an individual document.

Since that all the tweets are individual and have a maximum 140 characters for each tweet this method will be joining all the tweets together in a class. The result will be very long and unreadable, however this allows the model to look at the TF-IDF as a class-based perspective.

Then instead of the model takes into account the number of documents in TF-IDF it will take the number of classes instead, so all the changes results in the following formula:

Figure 3.1

$$W_{t,c} = tf_{t,c} \cdot \log(1 + \frac{A}{tf_t})$$

where the frequency of each word t is extracted for each class I and divided by the total number of words w. this action can be seen as a form of regularization of frequent words in the class. Next, the total, unjointed, number of documents m is divided by the total frequency of word t across all classes.

After applying all the previous steps, a model has been obtained and saved for future reference which has all the topics generated from the Spanish company, also this model can be applied on the sample data set that has been generated accompanied with the code. The sample data set was obtained from the original tweets for each account and have passed all the process as mentioned above for replication purposes.

RESULTS

3,879 topics where obtained from the Spainsh company with an outlier topic which consist of 79,925 words. On the other hand, a vegan topic was extracted and it was ranked the 6th with Word count of 668 with 202 users have been talking at least onces on the topic which represent 5.3 % of the original users .

In the following figure shows the most frequent words for each topic, our topic of interest is topic 5 which has all the words for vegan (Figure 4.1). Topics are ranked based on how many words are include in the Topic. Since Python counter from 0 so Vegan topic lands in the 6th place, we can see also the first topic is spainsh words that was not translated which is the highest count in topics followed by people talking about twitter, then comes in the 3rd, 4th, and 5th cities in spain and topics about spain, lastly comes the vegan topic which have a noticable high place at the ranking of the topics.

Figure 4.1

By matching the Topics from the spanish company to SBahn follwers tweets to identified that they have interest in vegan 4,457 users in the data set 123 of them was successfully matched to the topic of veganism which represnets 2.76% of the followers from SBahn users, from 420,762 tweet that was analysed from the account 199 tweets where flagged as veganism which represents 0.05 % .

LIMITATIONS

Twitter data often contains alot of noise that cant be treated such as mispelling and abbreviations that is caused by the limit of the length of the tweet, so the model could not understand the full contextual meaing of the tweet. Also BERTopic provide topic labels but does not offer explanations or interpretations of the topics. This limitation makes it challenging to understand the underlying meaning or connections between topics, which can be crucial for further analysis. On the other hand, the compuational power was a blocker since the tweets needed to be aligned on the same language to generate a valuable insight, so by not relying on cloud servers this method would take alot of time for the company, but if the company has access to a cloud server that they can utilize with high computational specfications it would be feasable, besides with the accessablity for a translation API the quality of translation could be higher. Also, data from Berlin was collected from only one account which can be not representable for people residing in Berlin.

DISCUSSION

Having 123 users from berlin that have interest in vegenism is not enough for the company to be able to take a decision, but comparably to the 202 users that was talking about vegenism from the Spanish company this accounts for 60% of the people which is a Good number but now sufficient enough for a company to explore the new market. A sample of the data has been produced to be able to replicate the study. Also there is a potential increase of the people who are interested in vegenism by analysing the followers and Friends of the users since followers and friends has the same intersets as the user and this reach can be expanded to just beyond the user (Adamic & Adar, 2003). With those users that are interested or talk about veganism if we calculated the 1% of their total followers it will be 3594.

By anaylsing other topics that was extracted from the spanish company their is a similarity between the topic of interest which is topic 5 and topic 7 which is about animal abusment that can also explain that people who talk about animal abusment may be also interested in veganism. This topic can be futher analysed to draw a better insight for the company expansion. Their is almost a similarity of 70% between the topics as shown in the similarity matrix (Figure 6.1),besides the hierarchical clustering also shows how topics are all related to each other and clearly topic 5 and topic 7 are clustered together (Figure 6.2).

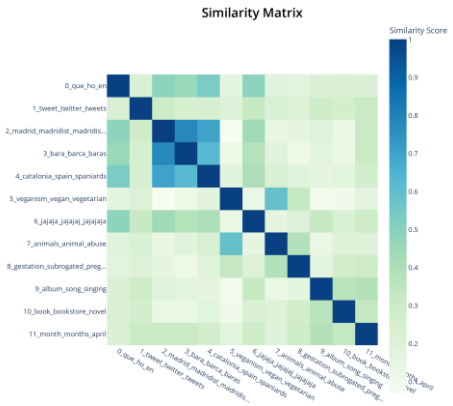

Figure 6.1

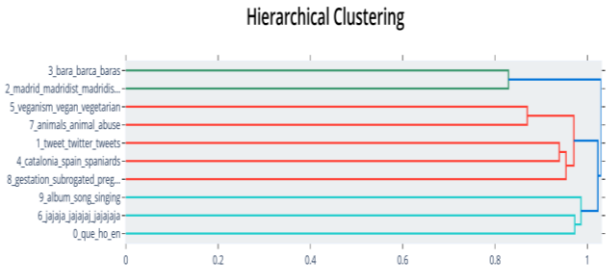

Figure 6.2

CONCLUSION

This paper aims to create a more feasable way for startups or SMEs to have a more efficient way to explore their market potential outside of their operating teritory, to achive this objective tweets were collected by Twitter API from two different acounts. The first account belongs to a company base in pain,while the second account is located in Germany. The company served as the foundation of the topic modeling using BERTopic which allowed the analysis of the target account which resides in Berlin, with the use if class-based TF-IDF. To prepare the text retrived from the API cetrain steps were taken to input the text to the model such as removing unneccesary text in the tweets and also encrypting the user related information. This alowed for the anonomity of the user. Corrolated topics was discovered through the analysis such as animal abusment was corrolated with veganism, which provide valubale insights for futher analysis. Lastly, the generated code of the paper is felxiable to use other data sets from twitter in respect to the column names and to be replicable for other companies and other cities, taking into account if there will be a translation step the correct model for the language should be used.

REFERNECE

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. Social Networks, 25(3), 211–230. https://doi.org/10.1016/s0378-8733(03)00009-1

Chung, H.-J., Chen, C.-C., & Hsieh, T.-J. (2007). First geographic expansion of startup firms: Initial size and entry timing effects. Journal of Business Research, 60(4), 388–395. https://doi.org/10.1016/j.jbusres.2006.10.021

Food Service Market Size, share & covid-19 impact analysis, by type (full service restaurants, quick service restaurants, institutes, and others), by service type (commercial and institutional), and Regional Forecast, 2022-2029. Food Service Market Size, Share, Trends | Growth Analysis 2029. (n.d.). https://www.fortunebusinessinsights.com/food-service-market-106277

Harris, L., & Rae, A. (2009). Social Networks: The Future of Marketing for Small Business. Journal of Business Strategy, 30(5), 24–31. https://doi.org/10.1108/02756660910987581

Jagongo, A., & Kinyua, C. (2013). The social media and entrepreneurship growth. International journal of humanities and social science, 3(10), 213-227.

Joachims, T. (1996). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. Carnegie-mellon univ pittsburgh pa dept of computer science.

Morrison, A., Breen, J., & Ali, S. (2003). Small business growth: Intention, ability, and opportunity. Journal of Small Business Management, 41(4), 417–425. https://doi.org/10.1111/1540-627x.00092

Plisson, J., Lavrac, N., & Mladenic, D. (2004, October). A rule based approach to word lemmatization. In Proceedings of IS (Vol. 3, pp. 83-86).

Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, No. 1, pp. 29-48)

state of the industry report Germany 2022. (n.d.). https://gfieurope.org/wp-content/uploads/2023/05/GFI-Europe-Sustainable-Proteins-in-Germany-Summary-EN.pdf

state of the industry report Spain 2022. (n.d.-b). https://gfieurope.org/wp-content/uploads/2023/04/2020-2022-Spain-retail-market-insights.pdf

Vegan food market size, share & covid-19 impact analysis, product type (vegan meat, vegan milk, and others), distribution channel (supermarkets/hypermarkets, convenience stores, online retails, and others), and Regional Forecast, 2021-2028. Vegan Food Market Size, Share and Growth Analysis [2028]. (n.d.). https://www.fortunebusinessinsights.com/vegan-food-market-106421