

PERCEPTION ABOUT GMOS BETWEEN THE USA AND EUROPE THROUGH TWITTER

A Paper
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By
Feng Chang

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

June 2017

Fargo, North Dakota

North Dakota State University
Graduate School

Title

PERCEPTION ABOUT GMOS BETWEEN THE USA AND EUROPE
THROUGH TWITTER

By

Feng Chang

The Supervisory Committee certifies that this *disquisition* complies with North Dakota
State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Kendall E. Nygard

Chair

Dr. Oksana Myronovych

Dr. Supavich Pengnate

Approved:

7/6/2017

Date

Dr. Kendall E. Nygard

Department Chair

ABSTRACT

Twitter is one of the most popular social networking and microblogging services in the internet world. By the end of Jan 24, 2017, there were at least 100 million active users daily around the world. Due to such a huge number of people, it is a great channel to collect information about almost anything existing in the world. Based on this information, we could also analyze the popular topics which are widely discussed. For the work of my paper, the target is to find out the perception about GMOs, but also focused on the difference in perceptions between Europe and the United States. To accomplish this work, a collection of Twitter was collected that all include the text about GMOs, along with their locations. Analytics were performed on the tweets to classify them by sentiment, then statistical tests were carried out to assess differences in perceptions by location.

ACKNOWLEDGEMENTS

I would like to give my sincere gratitude to my advisor, Dr. Nygard, for his valuable suggestions, patient guidance, and unceasing encouragement during my whole research project. I also would like to thank my supervisory committee, Dr. Myronovych and Dr. Pengnate. I really appreciate their time, kindness, and profound comments.

Also, I want to give my wife and family great thanks, for their accompanying, supports, encouragements, and love during my whole life.

DEDICATION

This paper is dedicated to my wife and parents for their endless love, support, and encouragement.

TABLE OF CONTENTS

ABSTRACT	iii
ACKNOWLEDGEMENTS	iv
DEDICATION	v
LIST OF TABLES	viii
LIST OF FIGURES	ix
1. INTRODUCTION	1
2. LITERATURE REVIEW	3
2.1. Background	3
2.1.1. Twitter Streaming API	3
2.1.2. Support Vector Machine Classification.....	3
2.1.3. Naïve Bayes Classification.....	4
2.1.4. TF – IDF	4
2.2. Related Work.....	5
2.3. Motivation	5
3. TWEETS DATA PREPARATION	7
3.1. Twitter Feeds Aggregation.....	7
3.1.1. Getting Authentication	7
3.1.2. Use Tweepy with Streaming API	8
3.1.3. Collect and Clean my Tweets.....	8
4. TWEETS DATA CLASSIFICATION	15
4.1. Support Vector Machine Classification	17
4.1.1. Set up Tools	17
4.1.2. Train Dataset	17
4.1.3. Test Accuracy	18

4.1.4. Classify Collected Data	19
4.2. Naïve Bayes Classification Process	20
4.2.1. Set up Tools	20
4.2.2. Train Dataset	20
4.2.3. Test Accuracy	20
4.2.4. Classify Collected Data	21
5. PROCESSED DATA ANALYSIS	23
5.1. Count Duplicate Words	23
5.2. Count Negative Word	24
5.3. TF – IDF Process	25
5.4. TF – IDF Results	25
6. CONCLUSION AND FUTURE WORK	28
6.1. Conclusion	28
6.2. Future Work	29
REFERENCES	30

LIST OF TABLES

<u>Table</u>	<u>Page</u>
1. 20 Lines Processed Tweets for the USA.....	12
2. 20 Lines Processed Tweets for London, England.....	14
3. Top 20 Words from txt files.....	24
4. Example of TF – IDF result.....	25
5. Top 10 Words in USA and Europe Results	27

LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
1. Create Your New Application.	7
2. Application Profile.....	8
3. Collecting and Cleaning Tweets.	9
4. A Sample of JSON Response.	10
5. The Structure of Classification Process.	16
6. The Size of Training and Testing data.....	17
7. The Size of Training and Testing data.....	18
8. Accuracy in SVM for different data.	18
9. Classified Tweets Data Percentage in SVM.	19
10. Accuracy in Naïve Bayes for different data.....	21
11. Classified Tweets Data Percentage in Naïve Bayes.	22
12. Result of Negative Words Count	24
13. TF – IDF Result for the USA.....	26
14. TF – IDF Result for Europe.....	27

1. INTRODUCTION

According to the 2015 revision of UN's project, the world population was expected to grow by 2.3 billion between 2009 and 2050. This means that we need to produce at least 70 percent more food production by 2050 for the population growth [1]. Therefore, to provide so much more food is an urgent task on the earth. We could ask people to stop wasting food, or create more land to farm, or come up ways to increase the yield. Genetically modified foods is one of the methods for increasing supply since genetic technology can boost the yield. Many believe that GMOs provide a very promising approach to feed the whole world, especially for those fast growing population countries.

However, we always say everything has two sides. Even though now we could use genetically modified foods technology to increase the yields and feed the whole world, we still cannot ignore its dark side. We wonder whether or not GMO foods are healthy or not. Nowadays, people often to share or express their ideas, opinions, events on those social media networks, especially on Twitter. The purpose of this paper is to collect Twitter feeds and work out analytics on the data to find out people's perceptions about GMOs. Two years ago, people from the same research team at NDSU had published a paper which concluded that regional locations in the United States would influence people's feeling toward GMOs [2]. Also, another researcher from the team wrote another paper on how gender differences [3] would affect people's perception in GMOs. Inspired by those two paper, this paper is to determine the attitude differences about GMOs between Europe and the United States.

To accomplish this goal, a Twitter – based analysis was formulated and applied. The study on Twitter is about collecting the tweets that use the keyword GMO. The tweets are then separated by different locations. The next major step is to classify how many tweets are positive

and how many are negative. After that, statistical calculation or tests were run to find out the different attitudes between the United States and Europe. Some people may choose to not reveal their locations in Twitter. t. That is a distraction for my research, so those tweets were removed.

The rest of the paper is structured as follows. The second chapter includes a literature review. The third chapter describes the way to capture and process the data. The fourth chapter describes the method to classify the data from the processed and cleaned data. In the fifth chapter, the classified data is analyzed, and some statistical results are provided for both United States and Europe. The sixth chapter is the conclusion of the whole paper and suggestions for the future work.

2. LITERATURE REVIEW

In this chapter, it will focus on covering the needed background, which is very important to the paper. Also, some similar research work will be discussed.

2.1. Background

2.1.1. Twitter Streaming API

Twitter provides their APIs to potential developers. Since mobile development is become a hotspot, more and more individual developer started to create their own applications or projects based on the services from Twitter, Facebook or Google. Also, many students and researchers started to use their services for research purposes. Therefore, Twitter allowed a developer to access their global streaming tweets data by using their Streaming API [4]. Twitter provided two APIs for the developer to collect tweets. One is the Search API, the other is the Streaming API. Based on two reasons the Streaming API was chosen. First, it is desirable that the tweets collecting process goes back in time. Second, a higher flow of tweets is available through streaming. The Streaming API delivers a maximum flow of 180,000 tweets per hour, which is two times more than Search API [5]. After Twitter data is collected, some cleaning work was carried out immediately.

2.1.2. Support Vector Machine Classification

Support Vector Machine (SVM) is a machine learning method which was established the 1990s. It partitions the data off into two sides based on a separating classification plane. Normally, a distance between a point or item and a classification plane measures the degree of accuracy. A SVM could maximize the distance, making it as accurate as possible. It enhances generalization ability by its Structural Risk Minimization. To put it simply, even though we might not have enough sample data, we still can achieve a good statistical result. Therefore, it

would be a good classification to use. The tool which this paper uses is scikit – learn’s LinearSVC. It’s because based on Marco’s experiment [6], LinearSVC is much faster than other SVC methods that are provided in scikit – learn. Chapter four will present more detail about this classification method.

2.1.3. Naïve Bayes Classification

Naïve Bayes Classification is based on Bayes theory. We have a formula [7]:

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

- $P(A|B)$: Probability of observing event A given B is true.
- $P(B|A)$: Probability of observing event B given A is true.
- $P(A)$: Probabilities of A
- $P(B)$: Probabilities of B

The Naïve Bayes ignores dependencies among events. This work uses basic Naïve Bayes tool which implements the same process in Python.

To implement this method, we have three steps. First, it requires classified data sample. Next, those classified datasets are provided to train the Naïve Bayes Classifier. Finally, a group of test data is needed to determine the accuracy of the trained Naïve Bayes Classifier. Chapter four will have the detail about the whole process.

2.1.4. TF – IDF

In TF – IDF, there are two terms [8]:

- Term Frequency is the ratio between the count of one word in a document and total word count in a document
- Inverse Document Frequency is the logarithm of the ratio between the total corpus and the corpus which has the specific word plus one.

Then, we multiple TF and IDF to get the result of TF – IDF. The bigger we got, the more the word is. Python NLTK package provides TF – IDF tool. Chapter Five will provide the detail about the information extraction by using TF – IDF.

2.2. Related Work

According to every customer's intention, we have a huge amount of research has been done which is related to GMOs. Also, we had always heard about that Europe was against GMOs the most. One article from The New York Times was written by Mark Lynas [9] attempted to tell us that Europe tries to avoid GMOs even turned against science. We could simply see that how serious Europe was to treated GMOs. They were not only saying that but also some countries like Germany, France, Greece, Italy, started to ban on the cultivation of genetically modified crops. Because Twitter is a good place for people to post their thoughts about something, researchers also start to capture and analyze Twitter feeds for multiple data mining purposes. Hanze, Li [10] examined the opinions and sentiments from his Twitter dataset on GMOs. In the paper, the attitude around the Europe or United States' on GMOs is the main target. Thus, both United States and Europe are two focused places, and their real difference is and what the real reason might be are the two main purpose needs to be revealed.

2.3. Motivation

Since GMO has been brought out, people start to debate on that. In 2013, two famous people in China had totally opposite ideas about GMOs and they were against each other. The

scientific organization like WHO (National World Health Organization) had stated that GMOs are safe, but still, lots of people don't believe that. They concern the safety, the generic pollution. On the other hand, grocery store provided labels to distinguish GMO and not – GMO food. This also made us consider if GMOs is safe, why they need labeling. Also, we just heard about people dislike GMOs, but do they or we really know what GMOs is? Or do people really dislike it or they just follow other's idea blindly? So many questions make me want to find out the truth behind everything.

A social network is a nice place for people to present their true feeling about something with others. Anytime you post something on Twitter, you will get plenty of replies which include strangers or friends. As time goes on, it became a big data collection which involves almost every topic. Therefore, it became the best dataset for researchers to do data mining things. It is fast to access, less limitation to follow, and more data field to get.

Furthermore, United States is one of the biggest developed countries. Europe is one of the biggest developed union. Therefore, their opinion about GMOs may reveal some real reason about people's attitude. Also, we had a stereotype that compares to the United States, Europe is stricter against GMOs. Based on less evidence or lack of the same type of paper online, the stereotype is doubted, so it is necessary to find out the truth.

3. TWEETS DATA PREPARATION

Even though there are lots of data mining research online, still less paper about GMOs attitude's comparison between the United States and Europe were found. Thus, this paper will make up the lack of the research. This research includes the complete process of collecting Twitter feeds, the whole process of cleaning the collected Twitter feeds, the datasets which are from other universities for classification training, the process of categorizing test data and further analysis.

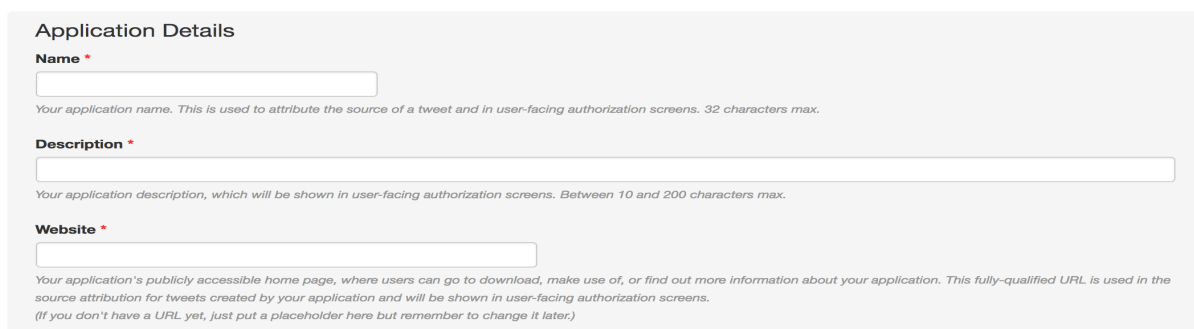
3.1. Twitter Feeds Aggregation

As I mentioned in the instruction, as an individual developer who wants to do a research or develop their own project, they have to work with Twitter API. However, everyone still needs to configure out the following steps.

3.1.1. Getting Authentication

Whenever you want to start your Twitter project, you will need a Twitter account. Then, you should go to Twitter application management website to create a new application based on filling out the required form which is shown in Figure 1.

Create an application



The image shows a screenshot of the 'Create an application' form on the Twitter developer website. The form is titled 'Application Details' and contains three main sections: 'Name', 'Description', and 'Website'. Each section has a text input field and a small red asterisk indicating it is required. Below each input field is a line of small text providing instructions. The 'Name' field is labeled 'Name *' and has a note: 'Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.' The 'Description' field is labeled 'Description *' and has a note: 'Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.' The 'Website' field is labeled 'Website *' and has a note: 'Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens. (If you don't have a URL yet, just put a placeholder here but remember to change it later.)'

Figure 1. Create Your New Application.

Once your application is created, you will have a profile page for your project like figure

forGMO

[Details](#)[Settings](#)[Keys and Access Tokens](#)[Permissions](#)

Figure 2. Application Profile.

Then, you go to Keys and Access Tokens and get consumer key, consumer secret, access token and access token secret. It means that every time when you use Twitter API, you must have those four keys in your application for credentials. If you miss any one of those or you misspell even one letter, your calls to Twitter will fail and you will get nothing.

3.1.2. Use Tweepy with Streaming API

Even though all the credentials are created in the account, the API connection is still not set up yet. The Tweepy package is the connection in Python between developers' call and API. It is an easy to use the library in Python for accessing Twitter API [11].

3.1.3. Collect and Clean my Tweets

Now, it is the data capturing process. Figure 3 provides the structure of my Twitter feeds collecting and cleaning process.

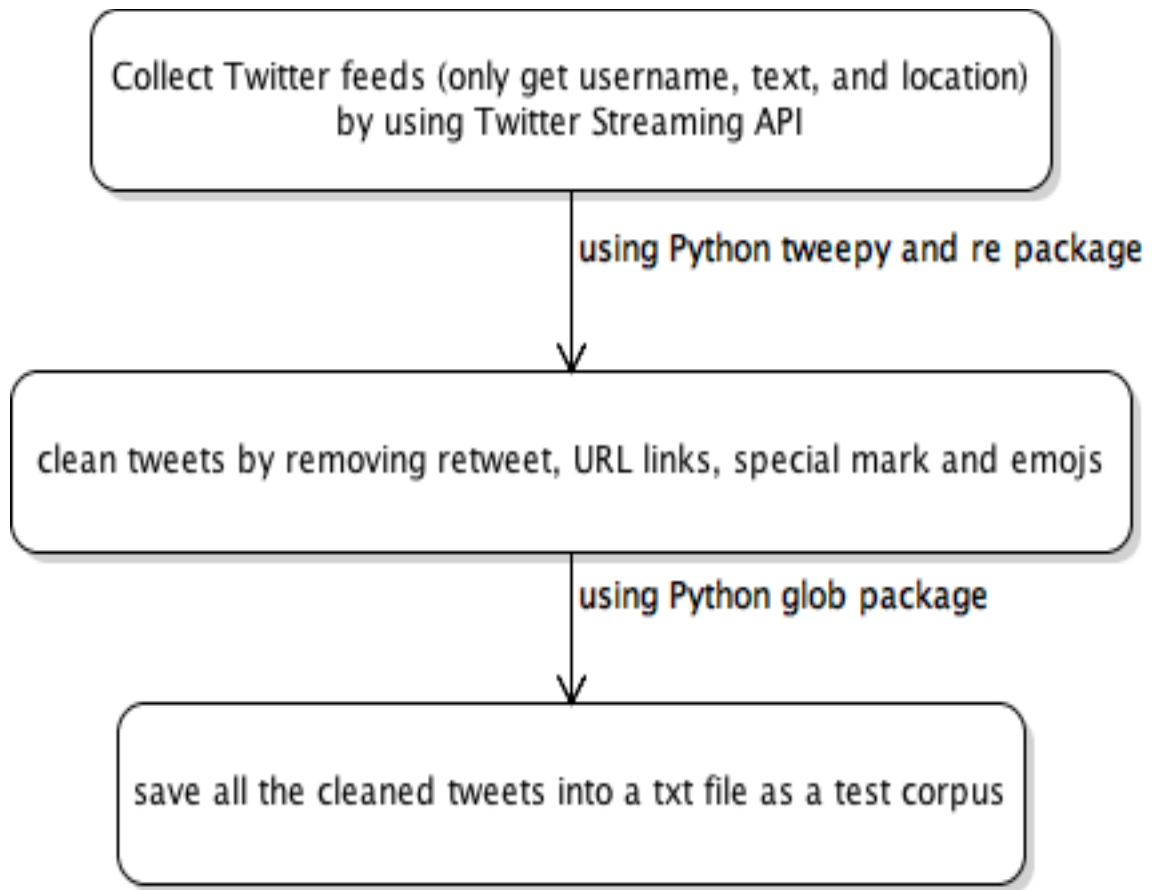


Figure 3. Collecting and Cleaning Tweets.

Note: For collecting and cleaning Tweets, it has three step totally. All of them are using Python and its tool package.

As the stated information above, this paper will focus on two things in tweets. One is location, the other is the text. However, for Twitter Streaming API, it will return me JSON type data like Figure 4 [12].

```

"contributors": null,
"retweeted": false,
"in_reply_to_user_id_str": null,
"place": null,
"retweet_count": 4,
"created_at": "Sun Apr 03 20:24:49 +0000 2011",
"user": {
  "notifications": null,
  "profile_use_background_image": true,
  "statuses_count": 31,
  "profile_background_color": "C0DEED",
  "followers_count": 3066,
  "profile_image_url": "http://a2.twimg.com/profile_images/1285770264/PGP_normal.jpg",
  "listed_count": 6,
  "profile_background_image_url": "http://a3.twimg.com/a/1301071706/images/themes/theme1/bg.png",
  "description": "",
  "screen_name": "PostGradProblem",
  "default_profile": true,
  "verified": false,
  "time_zone": null,
  "profile_text_color": "333333",
  "is_translator": false,
  "profile_sidebar_fill_color": "DDEEF6",
  "location": "",
  "id_str": "271572434",
  "default_profile_image": false,
  "profile_background_tile": false,
  "lang": "en",
  "friends_count": 21,
  "protected": false,
  "favourites_count": 0,
  "created_at": "Thu Mar 24 19:45:44 +0000 2011",
  "profile_link_color": "0084B4",
  "name": "PostGradProblems",

```

Figure 4. A Sample of JSON Response.

Note: The data returns from Twitter Streaming API has its JSON format and includes more than one or two tuple value.

First, Tweepy package in Python will extract username, text, and location for the United States. However, not every tweet has its location. Some people prefer to uncheck the option “show my location”. The tweets which don’t have location will be filtered out. Also, the method to locate a tweet is to set up by its location’s geocode. Second, when those tweets’ locations were identified, most of their text format is a city, state-based. Then, it is convenient to have an array

only includes all the states and plus an extra word, "USA" to match up with locations in tweets to filter out the US locations. This way has two benefits. One is to make sure tweets is from the US, the other is to make sure the location format stays the same. The final step, emojis, URLs, retweets or those special symbols need to get rid of from the tweets still. The tool is re package in Python. It only needs one simple line of code to remove all those things : “re.sub(r"(@[A-Za-z0-9]+)|([^\0-9A-Za-z \t])|(http\S+)|(RT @)","",item.text)”. Now the final processed US data is in Table 1.

Table 1
20 Lines Processed Tweets for the USA

Username	Text	Location
yeah_Im_gmo	Chad Reed is Elis last hope	USA
Samiam01x	Easy Ways to Go GMOFree and Why You Should gruccifer2 pjnet america	USA
NeoProgressive1	VirginiaInCal Boycott all Pepsi products like StacysPitaChips amp Sabra for blocking GMO labeling lawsTry Non GMO Verified brands in	KY
NeoProgressive1	VirginiaInCal Pepsi has dumped 12M into blocking GMO labelingBOYCOTT all Pepsi products like Cheetos for blocking GMO labeling la	KY
NeoProgressive1	VirginiaInCal Dont be tricked by Pepsi into buying beverages wtoxic GMOs like OceanSpray Boycott all Pepsi products for blockin	KY
city_market_CY	Mid night snack I ate the whole tub of guacamole with these Super clean Non GMO	TN
NeoProgressive1	VirginiaInCal PRODUCT WARNING Smuckers is selling unlabeled GMOs BOYCOTT all Smuckers products like Knotts 4 blocking GMO labeli	KY
jayizzo50	trutherbotred GMO Foods YOUR KILLER	CA
jlacy64	fuck gmos endGMOs	USA
randybear29	ImBeginningToSuspect depopulation started with GMOs	OH
tai_writes	xalimoos I cant wait to be a mom id pack my dogs lunch with organic no GMOs kibbles	USA
egojab	Jollett For most who oppose GMO its not really about safe but about natural	WA
Porscheey	Stop eating meat its bad for you GMO free	TX
DefendingBeef	Grass fedfinished cattle dont eat GMO corn or fillers Grasses dont require pesticidesSo just bei	CA
L_Gale517	ACOSorg jdaniel	USA
Taylor96Taylor	MeosoFunny Doesnt Eat GMO Because of Dangers Diet Consists of Drugs Alcohol and Trail Mix	USA
DTPORGE	NoGMOsVerified Select Committee report on GM is an insult to science and a danger to the public GMOs RightToKnow GMO gw	NY
DTPORGE	FarmFairyCrafts Boycott for Advertising in her Magazine GMO So Disappointed	NY
DTPORGE	FarmFairyCrafts via Second Silent Spring Bird Declines Linked to	NY

Table 1 shows 20 lines of processed Tweets which only have username, text, and location.

Next, a different process needs to be implemented for Europe Twitter feeds. To locate Europe, it is not so efficient to just use its geocode, so to pick up the main cities for Europe

countries is the best way. The reason is easy to find the matched geocode. Those cities are Paris, Berlin, London, Rome. People can see those cities are all main cities for their country, so they should be able to represent their country somehow. The rest of the steps for Europe tweets are the same. Therefore, Europe tweets data is in Table 2.

Totally, two months were spent to collect tweets. Next, one final step is to put them into one corpus for the next experimental stage. In this step, tweets were separated into the United States and Europe.

Table 2

20 Lines Processed Tweets for London, England.

Username	Text	Location
adonisfoods	Erythritol is 100 natural the one we use is organic and GMO free and come from fermented fruits	London, England
ashmakkar	I support bio	London
ashmakkar	India India Hindi	London
TheTapBlog	Dark History of Bayer Crop Science from manufacturing poison gas to hiding side effects of its min	London, England
Col_Connaughton	Lies Lies and More Lies GMOs Poisoned Agriculture and Toxic Scientific Rants GMO india monsanto	London UK
ActivistFangirl	HELP NEEDED Azure Organic Farm in Oregon will be forcibly mass poisoned with glyphosate by the county government	London, United Kingdom
LatestNewsOnDot	Author andrewcheetham Canadas Parliament To Vote On Mandatory Labeling For GMO Foods latestnews	London, England
weedseeds_UK	GMO weed	London, England
Col_Connaughton	July2315 False Flag Weekly News falseflag vaccination iran MH17 TTIP israel GMO fraud	London UK
buildeven	Modern agritech GMO needs a more lightly regulated US State to thrivebut its so not Iowa	London, England
jasminglynne	gmo is so fucked not from a health standpoint but from a corporate standpoint like wtaf	London
gmo_crops	GMO news Brazils Mato Grosso Leads Push for GMFree Soy The largest soyproducing state in Brazil Mato Gross	London
gmo_crops	GMO news Mustard Set to Be Indias First GM Food Gets Regulator Nod	London
gmo_crops	GMO research TLCUV hyphenated with MALDITOFMS for the screening of invertase substrates in plant extracts	London
gmo_crops	GMO research The use of Stationary Phase Optimized Selectivity Liquid Chromatography for the development of h	London
gmo_crops	GMO research Immunoaffinity chromatography combined with tandem mass spectrometry A new tool for the selectiv	London
vincentdignan	Hey GMO welcome Want to get your posts seen on Facebook Read this	London, England
pecasyrizos	Nothing less of organic gluten free gmo free raw vegan buckwheat brown bread with sprouted hemp seeds of course	London, England
bigpicturetv	The latest The Big Picture Daily Thanks to gmo mothersday	London
bigpicbiz	The latest The Peter Eyres Daily Thanks to tobias gmo sustainability	London

4. TWEETS DATA CLASSIFICATION

The main purpose is to compare tweets between USA and Europe to see the real attitude and to try to find out the real reason that Europe is against GMOs. Obviously, either way, to classify the sentimental results is the first duty. The data only has two polarities, positive and negative. It's because my training datasets are also having only two polarities. Support Vector Machine classifier will be trained first, then with Naïve Bayes Classification. Also, two datasets online were downloaded for training purpose. The complete structure of classification process is in figure 5.

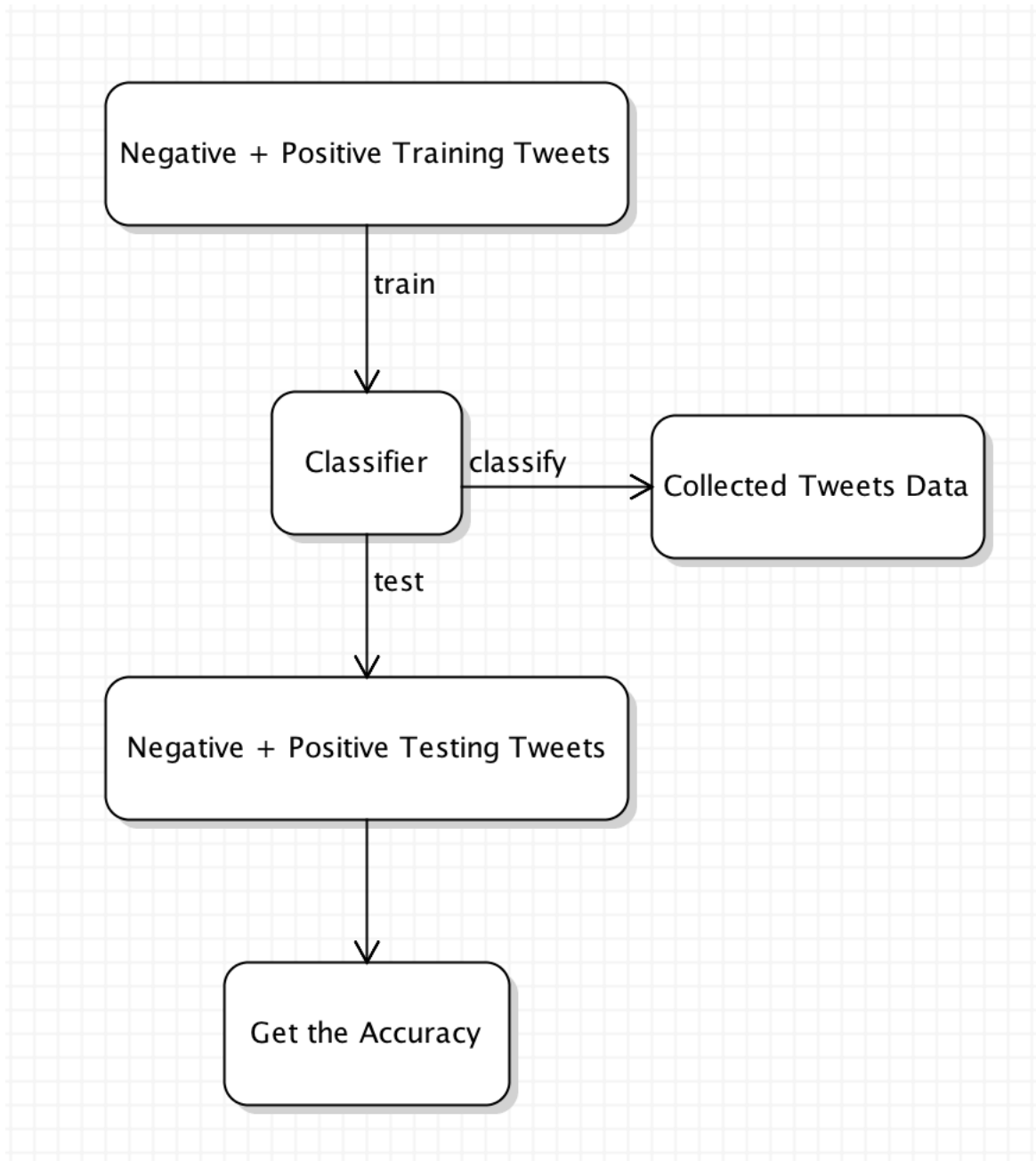


Figure 5. The Structure of Classification Process.

Note: The man – labeled Tweets dataset will be split into two part, training, and testing. Training dataset trains those classifiers for classifying collect data later. Testing dataset finds out the classifier's accuracy.

4.1. Support Vector Machine Classification

Based on chapter two, even though we might not have enough sample data, we still can result a good statistical regularity with SVM. Therefore, it would be a good classification to deal with the size of my data

4.1.1. Set up Tools

In Python, we have plenty of good tools to deal with sentimental analysis. The main two are NLTK [13] and Sklearn.svm. NLTK is a famous tool go work with human language data. It builds – in some easy to use methods for us to play with. First, every line in data needs to be separated into word format, so nltk.tokenize package help me with that. Next, if the classifier is well – trained, it will be better to test the data, so nltk.classify is chosen. Both of them are very handy. The rest leaves to Sklearn.svm. It provides LinearSVC for me to create my training model.

4.1.2. Train Dataset

The first dataset is from Bo Pang and Lilian Lee's [14] movie review data [15]. It included 5331 positive and 5331 negative snippets. The size of data separation is like figure 6.

```
training_data = pos_data[:4000] + neg_data[:4000]  
testing_data = pos_data[4000:] + neg_data[4000:]
```

Figure 6. The Size of Training and Testing Data.

Note: Training data has the first 4000 of 5331 from both positive and negative data. Testing data has the rest from both data.

The first 4000 data from both positive and negative data are used to organize a training data. Then, the rest of the data became the testing data. The ratio is about 3 approximately.

Next, it is a dataset from Alec Go, Richa Bhayani, and Lei Huang [16]. It is pure Twitter feeds data. And according to its size, only five thousand sentences are selected for each positive

and negative tweets. To match the previous ratio, the 3750 are for training and 1250 are for testing like figure 7.

```
training_data = pos_data[:3750] + neg_data[:3750]
testing_data = pos_data[3750:] + neg_data[3750:]
```

Figure 7. The Size of Training and Testing Data

4.1.3. Test Accuracy

Finally, figure 8 shows us two groups of quite high results.

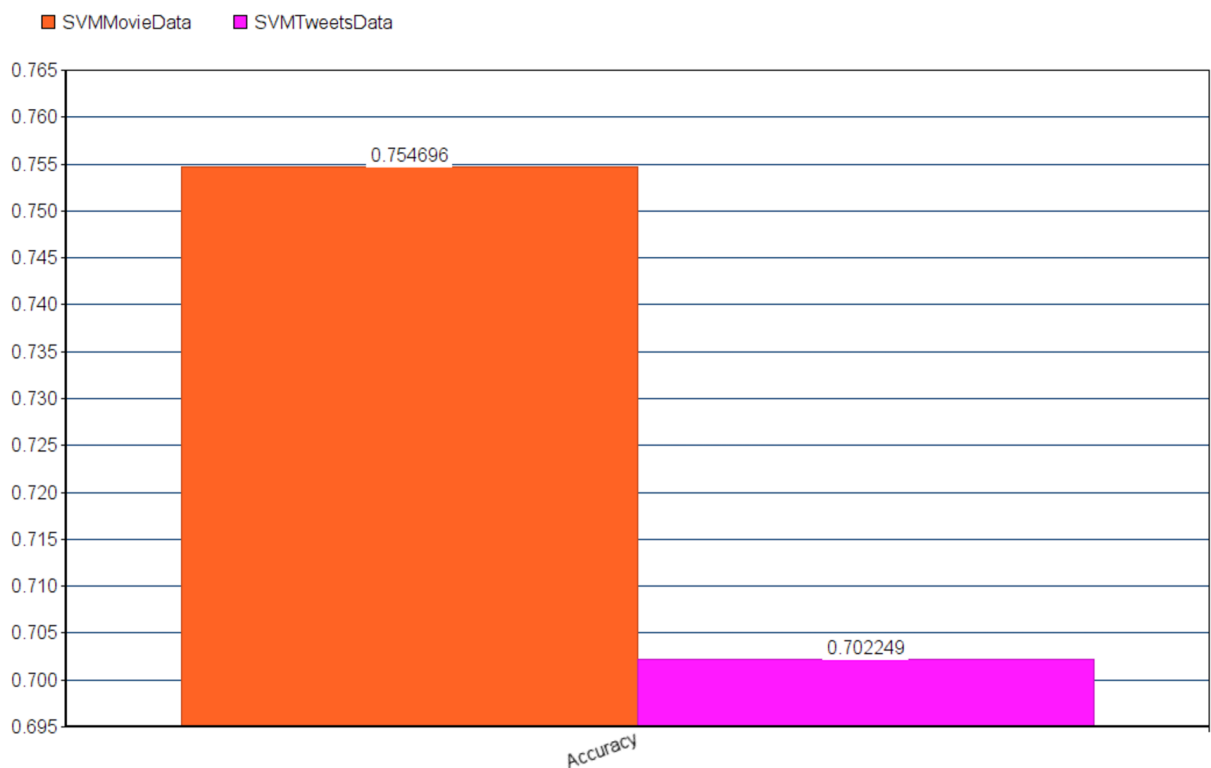


Figure 8. Accuracy in SVM for Different Data.

Note: In the SVM model, the accuracy is about 0.755 for movie data and is 0.702 for Tweets data. Even from the chart, the distance from pink top to the orange top is significant, it is 0.05 difference.

4.1.4. Classify Collected Data

Since the accuracy is so high on both training datasets, the research will use them to classify my own processed dataset. Mine is also pure tweets data includes 1097 positive and 1097 negative. This time a surprising result is shown in figure 9.

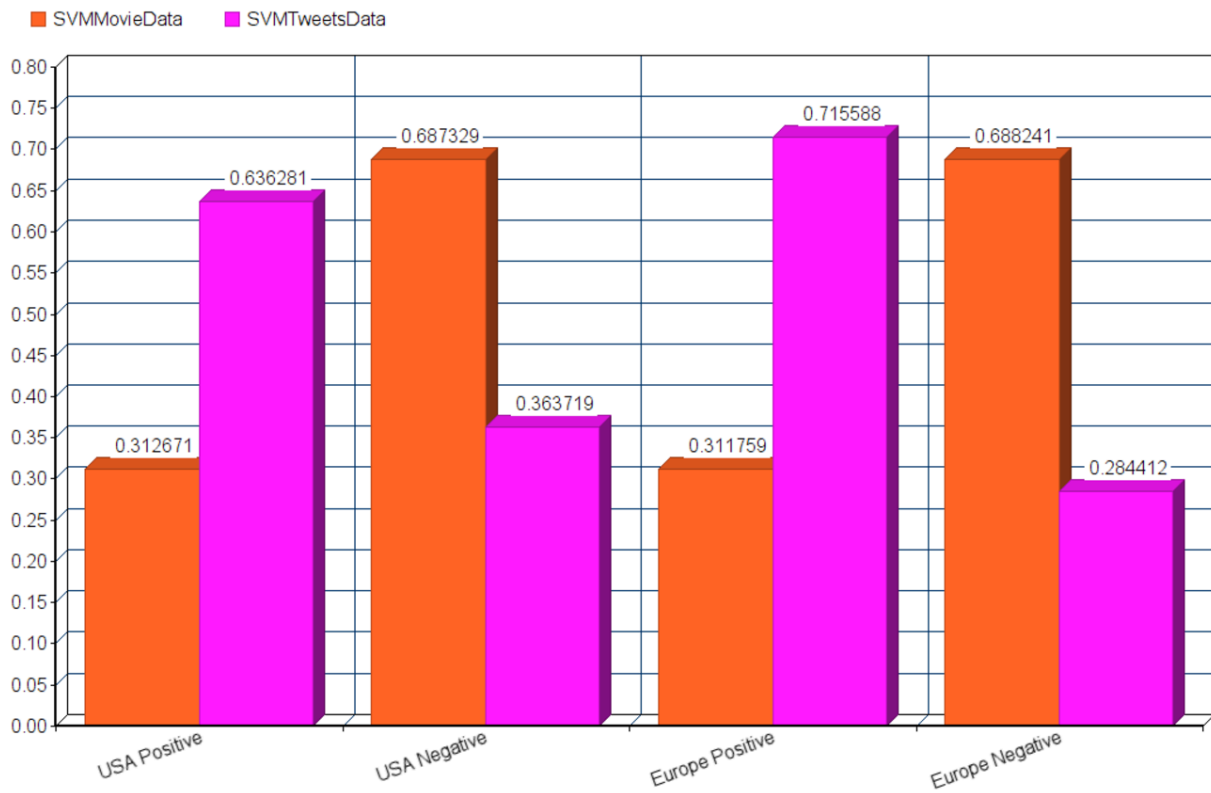


Figure 9. Classified Tweets Data Percentage in SVM.

Note: Support Vector Machine has about 0.687 for the USA negative and 0.688 for Europe negative in movie data, but it has about 0.364 for the USA negative and 0.264 for Europe negative in Tweets data.

From this chart, we saw a pair of opposite attitudes based on different data types. Chapter six will dig deep on that. For now, it will generate a txt file which stored those negative tweets from SVM model by using both datasets.

4.2. Naïve Bayes Classification Process

From book Data Mining Algorithms: Explained Using R [17], it stated that the naïve Bayes classifier is one of the simplest methods to achieve classification problem with a reasonable accuracy. Therefore, another way to test how well my dataset works is to use Naïve Bayes Classifier. And the main steps and datasets are the same.

4.2.1. Set up Tools

This time, the only tool is NLTK. It has everything built – in for Naïve Bayes. Those tweets still need to be separated into words format. Then, they will be trained and tested by Naïve Bayes Classifier. Next, the accuracy of trained model will be presented. At the end, the model will be used to classify the collected tweets.

4.2.2. Train Dataset

Due to the same standard, SVM used the same datasets. Also, the separation ratio is the same as figure 6 and 7.

4.2.3. Test Accuracy

For Naïve Bayes, the accuracy for movie data is still higher than tweets data. It's in figure 10. The value is still quite high.

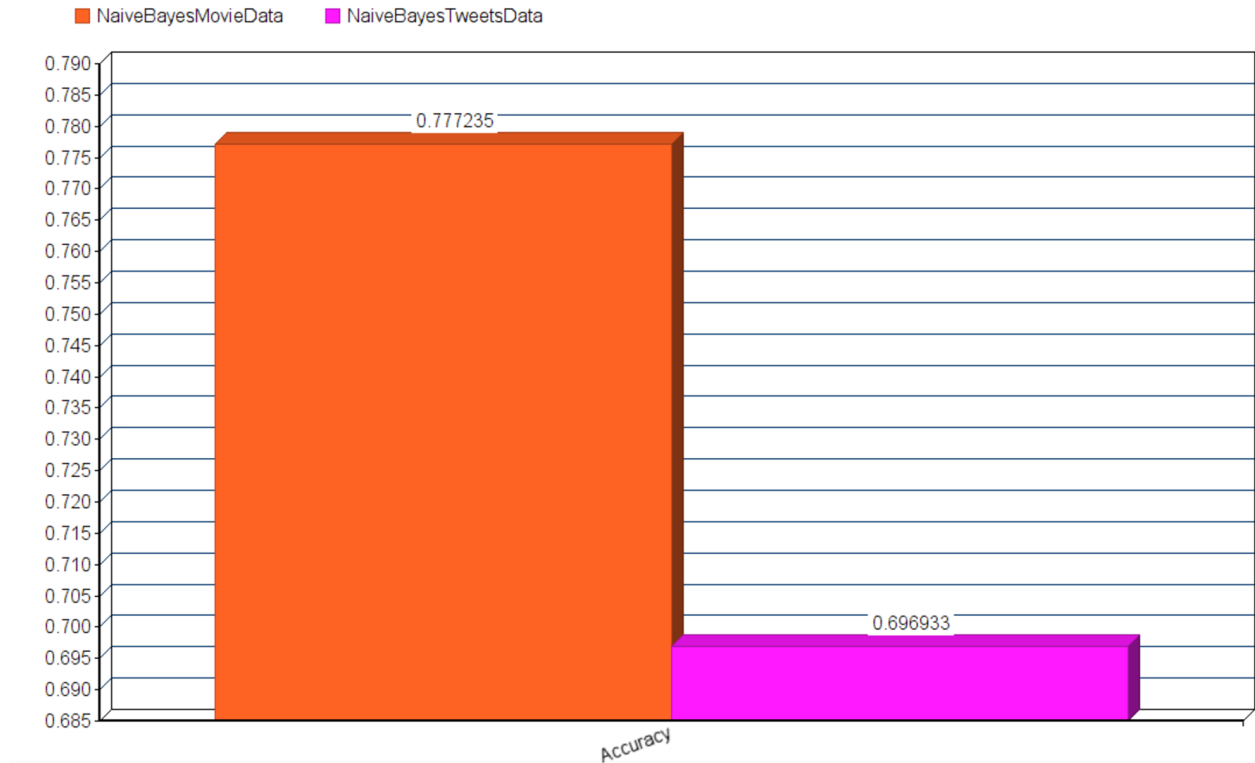


Figure 10. Accuracy in Naïve Bayes for Different Data.

Note: In the Naïve Bayes model, the accuracy is about 0.0.777 for Movie data and 0.697 for Tweets data. Even from the chart, the distance from pink top to the orange top is significant, it is 0.08 difference.

4.2.4. Classify Collected Data

Again, the trained Naïve Bayes model could be used to classify the collected tweets data.

The result is showing in figure 11.

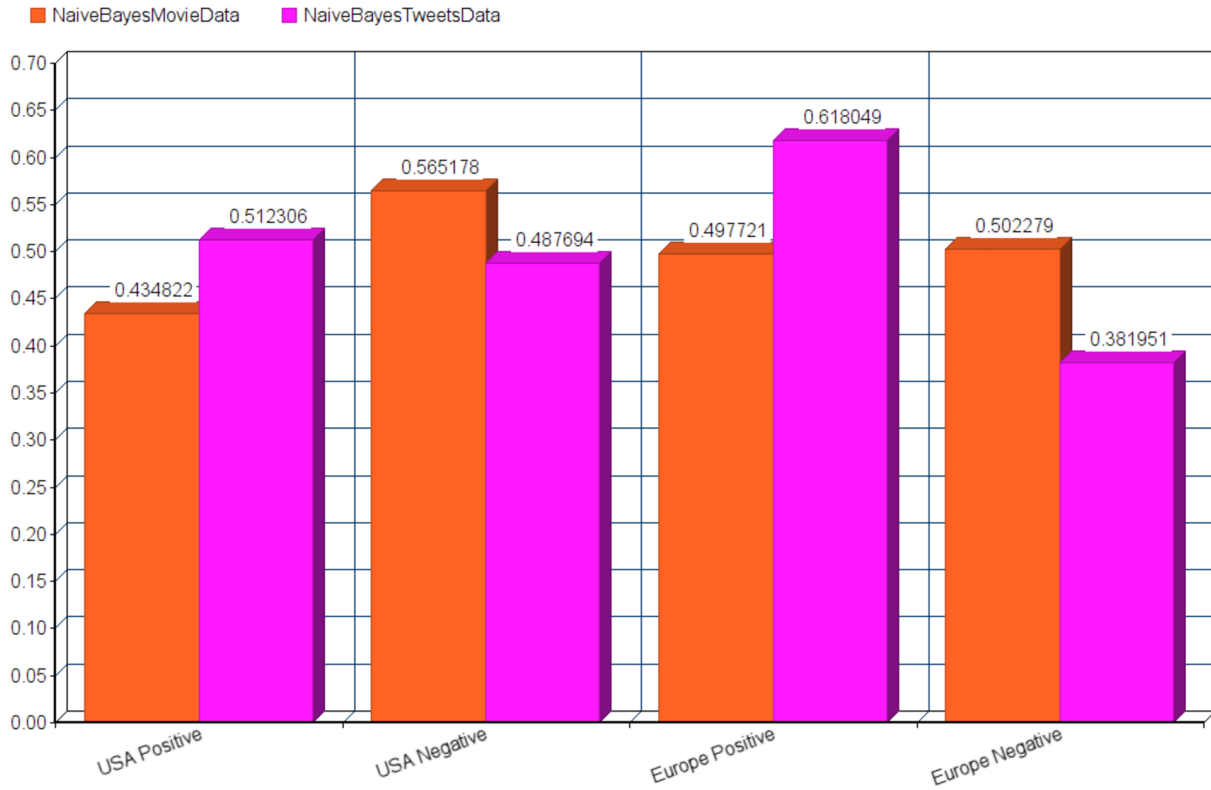


Figure 11. Classified Tweets Data Percentage in Naïve Bayes.

Note: Naïve Bayes has about 0.565 for the USA negative and 0.502 for Europe negative in movie data, but it has about 0.488 for the USA negative and 0.382 for Europe negative in Tweets data.

Overall, the Naïve Bayes results don't have any significant difference compared to SVM, but positive attitude increases in Naïve Bayes. Again, the paper will leave a discussion of that in chapter six and generate a txt file which stored those negative tweets from Naïve Bayes model by using both datasets.

5. PROCESSED DATA ANALYSIS

In the previous step, the SVM and Naïve Bayes models generate some txt files only include negative tweets. Those txt files will be analyzed to get a close answer of my question. Next, TF –IDF will be used here to extract the important information [18]. People often use it to retrieve information or mine text. It fits here because the reasons for denying GMOs may come from some keywords in those tweets or Twitter's text, so TF – IDF will help me analyze how important the specific word is a document in a corpus. Also, if the value of TF – IDF is bigger, the word is more important.

5.1. Count Duplicate Words

Before TF – IDF is implemented, the first method comes up in my mind is the simplest one, the most important word should appear the most in the corpus. Therefore, all the duplicate words are counted for each txt file. Based on all the results, table 3 to show a top 20 words which appears more than once from all negative tweets from Standford Twitter Data Naïve Bayes USA (TNBUSA), Standford Twitter Data Naïve Bayes Europe (TNBEURO), Movie Review Data Naïve Bayes EURO (MNBEURO), and Movie Review Data Naïve Bayes USA (MNBEUSA).

Table 3
Top 20 Words from txt files

TNBUSA	TNBEURO	MNBEURO	MNBUSA
gmo:406	gmo:342	gmo:434	gmo:474
the:214	to:147	the:244	the:250
to:187	the:126	to:213	to:227
and:125	of:90	of:115	of:131
a:119	in:90	in:110	a:121
in:117	news:87	a:84	in:116
of:110	and:78	and:81	and:103
is:109	a:64	on:80	is:100
i:86	is:63	for:73	i:83
for:69	for:51	news:70	gmos:83
gmos:68	i:41	is:68	you:81
that:62	on:36	i:56	for:78
on:58	plant:35	latest:43	on:76
food:56	with:35	via:43	are:73
are:54	from:34	plant:42	that:68
non:50	food:34	gmos:41	amp:57
it:47	no:32	research:39	it:55
monsanto:47	are:30	that:38	organic:54
not:46	not:30	daily:38	monsanto:52
with:45	that:29	you:38	corn:52

Table 3 show the top 20 duplicated words in all txt files.

5.2. Count Negative Word

Look at the table from the previous step, it is really hard to get any information from those most counted words. Then, to count only those negative words might be a better idea. To do so, I need to work with some English dictionary. I got a Harvard-IV_NegativeWordList [19]. This time a disappoint result shows in figure 12.

```
none of your words are listed in Harvard-IV_NegativeWordList.txt
>>> |
```

Figure 12. Result of Negative Words Count

No negative words were found in negative tweets which could match with the dictionary. It is easy to understand because users in Twitter doesn't type like we speak in daily life.

5.3. TF – IDF Process

In the end, TF – IDF might be the best tool. Compared to other two, it is more reliable because it could be used to get important words from each single sentence. For instance, the sentence, “Fashion tips Bowling Green is a very feminine textileTacrobes are not robes but GMO”, the TF – IDF tool will calculate a score or weight for each word in this sentence like table 4. Column Word is one tweet’s sentence word split. Column Document is the sentence position in the txt file. Column Score is the TF – IDF weight for each word in the same sentence.

Table 4
Example of TF – IDF result

Word	Document	Score
Fashion	0	0
tips	0	0
Bowling	0	0
Green	0	0
is	0	0
a	0	0
very	0	0
feminine	0	0.4
textileTacrobes	0	0.5
are	0	0
not	0	0
robes	0	0
but	0	0
GMO	0	0.2

Table 4 shows the possible example of TF – IDF result.

And, TF – IDF is also built – in Sklearn Python package.

5.4. TF – IDF Results

Combine all those txt files, two final results for the USA and Europe are showing. Figure 13 is for the USA, and figure 14 is for Europe. These are part of the TF – IDF results. Left is the word, right is the score.

```

(cocacola) --- (0.584198229092)
(blocking) --- (0.574128568488)
(boycott) --- (0.504644530205)
(labeling) --- (0.508172022058)
(potato) --- (0.519455295998)
(trial) --- (0.515488706209)
(hitting) --- (0.514297076966)
(4blocking) --- (0.567524365388)
(citation) --- (0.587660131385)
(boycott) --- (0.550837396536)
(contamination) --- (0.713413500011)
(nestle) --- (0.533995580938)
(retweet) --- (0.526826071649)
(blocking) --- (0.574128568488)
(boycott) --- (0.504644530205)
(labeling) --- (0.508172022058)
(scare) --- (0.72049412655)
(brain) --- (0.597306045584)
(damages) --- (0.597306045584)
(appreciated) --- (0.537114710764)
(gabbana) --- (0.537114710764)
(ingredient) --- (0.572249457097)
(number) --- (0.572249457097)
(avoid) --- (0.71008754904)
(foods) --- (0.54755107015)
(iowa) --- (0.510390216102)|
(shove) --- (0.501286553804)
(straightforward) --- (0.503227576823)
(awesome) --- (0.56087237307)
(defining) --- (0.56087237307)
(explanation) --- (0.56087237307)
(charm) --- (0.61768794969)
(blocknig) --- (0.710212898246)
(veres) --- (0.585943083403)
(try) --- (0.526812363137)
(toxinfree) --- (0.65728518573)

```

Figure 13. TF – IDF Result for the USA.

Note: After implementing TF – IDF, the result is showing the important word which has the value higher than 0.45 for the USA.

```

(league) --- (0.492740366202)
(m8) --- (0.492740366202)
(rocket) --- (0.492740366202)
(bought) --- (0.492815185095)
(pp) --- (0.514760869635)
(touch) --- (0.514760869635)
(already) --- (0.491793268657)
(2020) --- (0.554339727874)
(keem) --- (0.587543065751)
(president) --- (0.53078158474)
(guys) --- (0.700850467811)
(love) --- (0.576199889492)
(exclusives) --- (0.581877732741)
(hype) --- (0.606650822187)
(still) --- (0.513760216327)
(waiting) --- (0.606650822187)
(titanfall) --- (0.535088542587)
(black) --- (0.512250741948)
(ops) --- (0.512250741948)
(bf4) --- (0.520650049738)
(matches) --- (0.520650049738)
(worst) --- (0.491226981839)
(cuh) --- (0.558094978503)
(invited) --- (0.526555816153)
(swayy) --- (0.558094978503)
(accepted) --- (0.726601953948)
(get) --- (0.504086815906)
(pass) --- (0.49723389046)
(westie) --- (0.527016754719)

```

Figure 14. TF – IDF Result for Europe.

Note: After implementing TF – IDF, the result is showing the important word which has the value higher than 0.45 for Europe.

Also, after sorting the values in the results, the top 10 word for bother results are showing in Table 5.

Table 5
Top 10 Words in USA and Europe Results

Top 10 USA	Top 10 Europe
labeling	facts
la	fears
twizzler	safety
hershey	foods
for	and
boycotting	of
reeses	the
stand	gmo
virginiaincal	explain
laws	explained

Table 5 shows the top 10 words in TF – IDF results for both USA and Europe.

6. CONCLUSION AND FUTURE WORK

In this chapter, it is an overall review of this paper and also discusses how we can improve this study in the future.

6.1. Conclusion

Overall, for this paper, the purpose is to find what the real attitudes on GMOs are in the USA and Europe and the truth behind people dislike GMOs. All the necessary steps are implemented based on the designed structure.

First, it created a Twitter account for authentication purpose, so that people can get all Twitter API keys include API key, API secret, Access token secret and Access token. Next, those four credentials must be added into Python code to connect Twitter Streaming API with the Python code to collect new dataset. After that, all those easy to use, built-in Python tool helped to extract information from tweets and clean those tweets. Also, SVM and Naïve Bayes are trained by using some good online datasets to classify the new dataset. Once the collected tweets are separated into positive and negative. Then, based on TF – IDF, it presented some clue which might provide some useful information in the future research. And, it got lots of interesting results finally.

In the end, the experimental results gave some answers for most of the questions. Starts now, people would not say they are 100% sure Europe is stricter to GMOs. One of those figures in Chapter four shows that the negativity is higher in the USA. Also, a thought might be true, sometimes people may just follow other's ideas, like against GMOs. If people really know why they hate it, when count negative word is counted in the dataset, it definitely should have some match results. However, the last doubt still hard to find out, it is the truth behind against GMOs which also brings out the future work for next part.

6.2. Future Work

I want to add a new thing for future work first. This paper is using Streaming API with string "gmo", but the return the Tweets are never checked. It might not be related to GMOs. Therefore, it is important to check if the Tweets is really about GMOs. Like previous part mentioned above, there is no answer for the truth behind people against GMOs yet. To look at those important words from TF – IDF, it is really hard to get any idea from that, but still, people could see some words like safety, poisonous, healthy, etc. Also, every time when N – gram is added into TF – IDF to assist in getting a better result, the laptop just slows down and not responds for a long time. Therefore, in the future, researchers could use a more powerful computer and combine with more precise language skills to dig deep into those words. Then, they could categorize them, so that they might have some ideas about the real reason that people hate GMOs.

REFERENCES

- [1] HIGH-LEVEL EXPERT FORUM. (2009). How to feed the world in 2050: ‘Global agriculture towards 2050’, at World Summit on Food Security, Rome, Italy.
- [2] Dass, P., Chowdhury, M., Lampl, D., Nygard, K. (2015). Risk Perceptions for Genetically Modified Organisms: An Empirical Investigation, at the 34th IASTED International Conference, Marina del Rey, USA.
- [3] Lu, Y. (2016). *Perceptions of Genetically Modified Foods by Gender* (Master’s Paper, North Dakota State University). Retrieved from <https://library.ndsu.edu/ir/handle/10365/25844>
- [4] Twitter Developer Document. (n.d.). Streaming APIs. Retrieved from <https://dev.twitter.com/streaming/overview> (Accessed: 2016-8-30)
- [5] 140DEV. (n.d.). Aggregating tweets: Search API vs. Streaming API. Retrieved from <http://140dev.com/twitter-api-programming-tutorials/aggregating-tweets-search-api-vs-streaming-api/> (Accessed: 2016-9-13)
- [6] Bonzanini, M. (2015). Sentiment Analysis with Python and scikit – learn. Retrieved from <https://marcobonzanini.com/2015/01/19/sentiment-analysis-with-python-and-scikit-learn/>
- [7] Saed, S. (n.d.). Naïve Bayesian. Retrieved from http://www.saedsayad.com/naive_bayesian.htm
- [8] XueMing, L., HaiRui, Li., Xue, L., & GuangJun, He. (2012). TFIDF algorithm based on information gain and information entropy. *Computer Engineering* Volume 38, 37-40.
- [9] Mark, L. (2015, Oct. 24). With G.M.O. Policies, Europe Turns Against Science. *The New York Times*. Retrieved from https://www.nytimes.com/2015/10/25/opinion/sunday/with-gmo-policies-europe-turns-against-science.html?ref=todayspaper&_r=0

- [10] Li, H. (2016). *Sentiment Analysis and Opinion Mining on Twitter with GMO Keyword* (Master's Paper, North Dakota State University). Retrieved from <https://library.ndsu.edu/ir/handle/10365/25787>.
- [11] TWEETPY. (n.d.). An easy-to-use Python library for accessing the Twitter API. Retrieved from <http://www.tweepy.org/>
- [12] HRP. (n.d.). Example JSON response from Twitter streaming API. Retrieved from <https://gist.github.com/hrp/900964>
- [13] NLTK. (n.d.). Natural Language Toolkit. Retrieved from <http://www.nltk.org/>
- [14] Pang, B., & Lee, L. (n.d.). Movie Review Data. Retrieved from <http://www.cs.cornell.edu/people/pabo/movie-review-data/>
- [15] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*-Volume 10(pp. 79-86). Association for Computational Linguistics.
- [16] SENTIMENT140. (n.d.). Stanford link. Retrieved from <http://help.sentiment140.com/for-students/>
- [17] Cichosz, P. (2015) Naïve Bayes classifier, in *Data Mining Algorithms: Explained Using R*, John Wiley & Sons, Ltd, Chichester, UK. doi: 10.1002/9781118950951.ch4
- [18] Chowdhury, G. G. (2010). *Introduction to modern information retrieval*. Facet publishing.
- [19] Joseph, P. (2012). Sentiment Analysis Datasets. Harvard IV_Negative Word List_Inf.txt. Retrieved from https://github.com/jperla/sentiment-data/blob/master/harvard_negative/Harvard%20IV_Negative%20Word%20List_Inf.txt