PERCEPTIONS OF GENETICALLY MODIFIED FOODS BY GENDER

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

By

Yang Lu

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

Major Department:
Computer Science

June 2016

Fargo, North Dakota

North Dakota State University
Graduate School

Title

PERCEPTIONS OF GENETICALLY MODIFIED FOODS BY GENDER

By

Yang Lu

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. Gursimran Walia

Dr. Limin Zhang

Approved:

9/27/2016	Dr. Brian M. Slator
Date	Department Chair

ABSTRACT

Twitter is one of the most popular worldwide social networking services. It has more than 320 million monthly active users around the world. So it's a very good way to discover what's happening in the world and we can even get people's opinions of some topics through their posts. Genetically modified food is one of the hottest topics all over the world. For the work of this paper, our aim is to determine people's opinions concerning GMOs, but also interested in whether there are differences by gender. To achieve the goal, the idea is to capture a large set of Twitter feeds that all include a reference to GMOs, then carry out analytics on the tweets to classify them by gender, then carry out statistical tests aimed at identifying differences in perceptions by gender.

ACKNOWLEDGEMENTS

At very first, I would like to express my sincere gratitude to my advisor, Dr. Nygard, for his valuable suggestions, wonderful guidance, and ceaseless encouragement during the research project. I also would like to thank Dr. Walia and Dr. Zhang for taking out time to be members of my supervisory committee. I really appreciate their time, kindness, and insightful comments.

Also, I want to thank my dearest family, for their encouragements, supports, and love throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

According to the latest UN projects, world population is expected to grow 2.3 billion between 2009 and 2050. This implies that we need to grow 70 percent more food by 2050 [1]. So feeding a third more mouths will be a big problem for the whole world. We can try to chop down forest to create ground for drops, which few people wants to do, or develop ways to boosting crop yields. Genetically modified foods (GMOs) could help to meet the demand. Some people believe that it will be impossible to feed the growing population of the world without genetically modified food (crops).

However, just as a coin has two sides, when we discuss about how the genetically modified foods can increase the crop yield and help to feed the whole world, a frequently asked question whether or not genetically modified foods are safe. It is clear that millions of people openly express their opinions concerning GM food on social media networks, especially Twitter. So the idea for this study is to capture Twitter feeds and carry out analytics on the data to help understand people's perception of GMOs. Earlier published research at NDSU concluded that regional locations play a role in forming people's perceptions toward GMOs [2]. Inspired by this research, we arrive at the research question concerning the possibility of gender differences play a role in people's perception in GMOs.

To accomplish this goal, members of our research group formulated and applied a survey-based method. The research of this paper follows that method with Twitter-based analyses. The Twitter study is based upon a captured set of twitter feeds that use the keyword GMO. The tweets are then categorized by gender using an independent python package. Enhancements to the package provided unprecedented accuracy in classifying gender. The next

major step is to extract the positive and negative words from each tweet and run some statistical tests to determine the association between gender and the perceptions expressed in each tweet.

In certain cases, the same person posted several tweets that were classified in a mix of male and female, which means, this is a false positive. To reduce the number of false positives, a voting scheme was devised. Basically, the scheme identifies any tweeter who posted more than one tweets and their corresponding gender. The tweeter's gender is then specified by a majority vote.

The remainder of the paper is organized as follows. The second chapter contains a literature review. Chapter three describes how the data was captured from Twitter and the gender of the tweeter was classified. Chapter four describes what the voting scheme is and how it helps with improving the accuracy of gender identification. In the next chapter, we provide statistical results and how gender differences affects on people's perceptions of GMOs. Chapter six contains the conclusion and description of future work.

2.   LITERATURE REVIEW

In this chapter, it is devoted to covering the necessary background, which is important to the study. Moreover, I overview relevant research as a comparison with my work.

2.1. Background

2.1.1. Twitter Streaming API

The Streaming APIs create a connection for developers to access Twitter's global stream of Tweet data [3]. Actually, Twitter provides Streaming API and Search (rest) API. The very essential difference is that Search API goes back in time but Streaming API goes forward. Another one is Streaming API returns a much higher flow of tweets than the Search. The Streaming API could deliver up to a flow of 180,000 tweets an hour. The Search API returns up to 72,000 tweets per hour. Since in the study, the priority is to capture a large set of tweets so we decided to use Streaming API to capture tweets by setting the keyword as GMO. After I collected all the data from Twitter I picked two main fields, which are username and tweet to do the analysis.

2.1.2. Genderizer

Genderizer is a language-independent module, which tries to detect gender information by looking at first name and/or making text analysis [4]. In the Genderizer system, naïve Bayes classification is used. Before we get started to use the Genderizer package, we should install it on our computer by running the command: "sudo pip install genderizer". Since Twitter does not provide user's real gender so Genderizer is a very good tool to help us to infer users' gender for twitter users. Chapter three shows more detail about Genderizer package.

2.1.3. Naïve Bayes Classification

Naïve Bayes classifier is based on applying Bayes' theorem as follow:

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

- 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 idea behind this is trying to compute the probability of being a male or a female by mining the text. As a supervised learning method, Naïve Bayes classifier often performs very well in practice. Without a question it's also very simple, so if we are looking for some pretty straightforward and well performance method then Naïve Bayes classifier is a good option. In the Genderizer package, Naïve Bayes classifier is used and Naïve Bayes classification is adopted.

2.1.4. Chi-square Test and P Value

Here is the definition of Chi-square: it is a statistical test commonly used to compare observed data with data we would expect to get according to a specific hypothesis [5]. It can determine whether there is a significant relationship between two categorical variables, in our case they are gender and the nature of the tweets.

The P-value is used to determining "unlikely" or "likely" by determining the probability [6].

- If P< 0.05, we can infer to statistically significant. 0.05 is often used as a cutoff between significant and not-significant results.

- If P<0.001, we can infer to statistically high significant.

4

2.2. Related Work

There is a considerable amount of research has been done on the GMOs field. One paper was written by Montserrat Costa-Font, Jose´ M. Gil, and W. Bruce Traill [7] attempted to bring evidence from several studies together and tried to find out consumer attitudes to GM food whether under some circumstances consumers willing to have GM food or they are willing to pay some more for Non-GM food. There are also some studies involving capturing Twitter feeds and doing some analysis based on those tweets. Agarwal and Sabharwal [8] extracted and analyzed a single tweet and followers and followed by some sentiment analysis. In my study, besides capturing tweets from Twitter also I extracted relevant keywords from individual tweets, after that, I ran some analyses directed at understanding those tweeters' opinions. More details are provided in the later chapters.

2.3. Motivation

As I expressed before, GMOs is a very hot topic now all across the world. Several years ago, most of the world didn't even have any idea what GMOs were, but now it seems most people know what GMOs are even my 91 years old grandma can tell me how she concerns the safety of GMOs. A funny thing is even some organizations within a scientific community like WHO (National World Health Organization) have stated that GMOs are safe but still so many people doubt about the conclusion. We most likely see some products with a sticker showing the product is Non-GMO, when we walk down grocery store aisles. That's the most important reason why we are so interested in people's perceptions of GMO.

The social network is a very good way for people to share their own ideas with others. Meanwhile, it's easy for us to get more opinions from those social network people. The best profit is we can collect Twitter data from people who come from different regions, have different

ages, ethnicity, and backgrounds. It can make our data more completed and comprehensive. Somehow, it's better to get data online rather than do some traditional surveys because people might show their ideas indirectly when they are doing face-to-face surveys.

Moreover, the gender differences also grab our attentions. Firstly, we think females might go to grocery stores purchase food often than males, especially for mothers. So we want to know if females are more concern GMOs safety problems than males, or it's totally oppisite. Another thing is GMO problem itself more looks like a science problem. According to some studies and statistics, only a few females study in STEM fields compare to males or we can say males might have more interesting and chances to read some science reports relate to GMOs. Since most of the studies show that GMOs are safe so we are very interested in if those statistics can affect them more positive for GMOs or it doesn't do any help.

Based on all the previous reasons, our group did this research and published the *paper Gender Difference in Perceptions of Genetically Modified Foods* [9]. The 31th International Conference on Computers and Their Applications (CATA) conference have accepted it. In that paper, I'm leading the Twitter analysis part and I'm able to show more detail in the following chapters.

3. RESEARCH APPROACHES

Even some studies have been done based on Twitter analyses, but I don't see any work have been done on evaluating differences in gender view on GMOs rely on Twitter analyses. So this chapter presents some more details about what we've done for the twitter study. It describes the process of capturing Twitter feeds, identifying gender for each user, parsing dictionary, tallying positive and negative words in tweet, and further analyzed. Figure 1 provides the big picture of Tweet analysis processing.
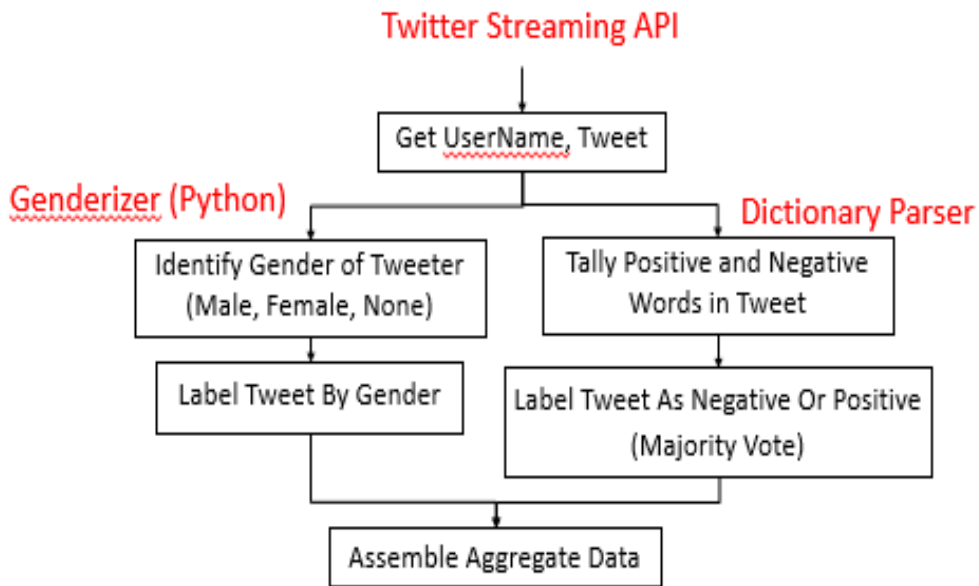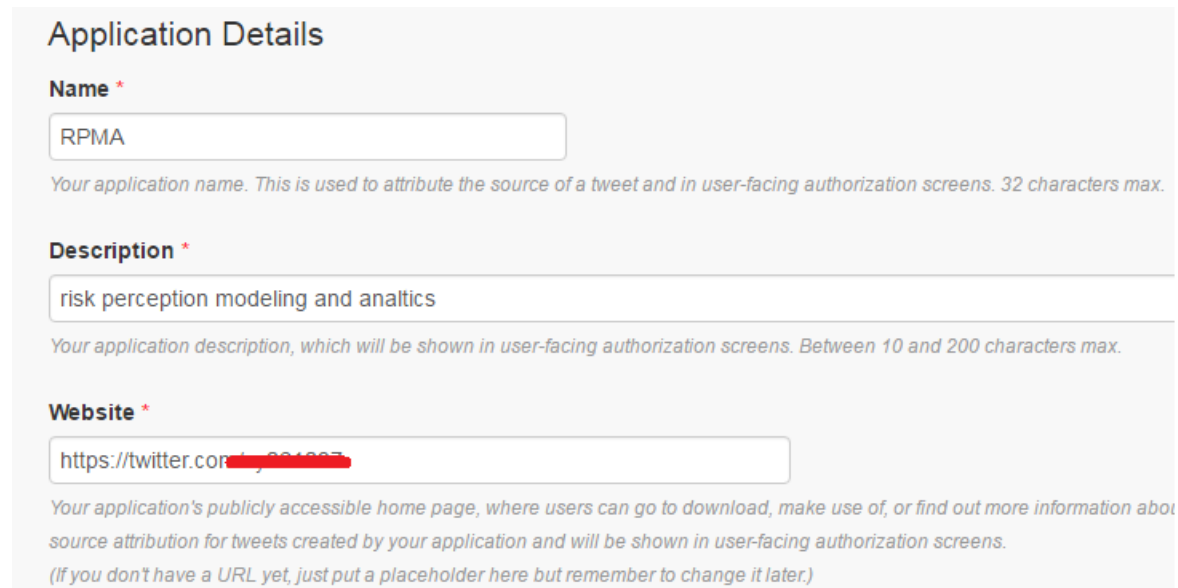

Figure 1. Processing Tweets

3.1. Twitter Feeds Aggregation

In order to use Streaming API to capture tweets related to keyword: GMO, we need to go through the following steps first.

3.1.1. Getting Twitter API Keys

Firstly, we need to create a twitter account if we don't have one. Then we should visit twitter application management website to create a new App and fill out the following form.

Figure 2. Create a New App

After the twitter app has been created, then we can get four keys. "API key", "API secret", "Access token", and "Access token secret". The following figure shows what the keys are. Those four keys contain the user credentials to access Twitter API.

Figure 3. Get API Keys

3.1.2. Start to Extract Tweets

After successfully got those keys, then we need to install a Python library Tweepy to connect to Streaming API. Now, everything is ready and we can write Python code to capture tweets from Twitter.

3.2. Gender Identification

The aim of this study tries to find out whether and how gender differences affect on people's perceptions of GMO. However, for some reasons we are not able to get user's real gender directly from Twitter. Then I did some online research and found some tools we could use. For example: Generize.io [10], Genderizer table [11], and so on. After some comprehensive consideration, I decided to use Genderizer as a tool in this study based on the following reasons.

Firstly, it's totally free. Secondly, Twitter is a widely used social network around the world so users could be anywhere in the world and people use different languages, so username could be an English name, Turkish name, Chinese name, and so on so forth. Genderizer supports different languages then we don't need to worry about the language problem. Moreover, as we

9

know sometimes we can't identify gender only by a username because people could set whatever they want as a username and it could be meaningless for example: 128824kn, 4869_hide, and so on. Genderizer also provides a way to figure out the problem; even the accuracy is not 100%. What does Genderizer do is, it makes text analysis using a naïve Bayes classifier [12] if the username does not infer any gender for sure. By using Genderizer module we can label user's gender as a male, a female, or none.

Table 1. Examples of Genderizer Classification

UserName	Tweet	Genderizer Classification
_andrewprlor	Make sure they're non-GMO, healthy goodies.	female
_bhickman	Thanks so much to GMO Free USA -- not just for helping to spread the word about the book, but for all of the...	male
_el_burro	I agree! Heard from a few people in Mexico Expressing concern over this change in gmo ruling.	None

3.3. Dictionary Parser

It contains a list of 2006 positive words and 4783 negative words from a KDD-2004 paper [13]. I extracted both positive and negative words from each tweet by matching the context in each single tweet with the word list. After this step, I got a total 520 positive words and 364 negative words for all the tweets.

Table 2. Examples of Extracting Positive and Negative Words from Each Tweet

Tweet	Positive Words	Negative Words
Make sure they're non-GMO, healthy goodies.	healthy	
Thanks so much to GMO Free USA -- not just for helping to spread the word about the book, but for all of the...	free, helping	
Pesticides in paradise: Hawaii's spike in birth defects puts focus on GM crops		defects
Please share: We are #Moms4GMOs! Scientist and Advocate Moms to Celeb Moms: Weigh GMO Food with Facts Not Fear	advocate	Fear, garbage

3.4. Data Assembling

After I got both gender identification and positive, negative words for each tweet, I tried

to split tweets by gender. I have a list contains all male users' names and their corresponding

tweets with extracted positive, negative words.

Table 3. Example of Processed Tweets

UserName	Tweet	Genderizer Classification	Positive Words	Negative Words
_andrewprlor	Make sure they're non-GMO, healthy goodies.	female	healthy	
_bhickman	Thanks so much to GMO Free USA -- not just for helping to spread the word about the book, but for all of the...	male	free, helping	
_bias	Pesticides in paradise: Hawaii's spike in birth defects puts focus on GM crops	male		defects
beer_amy	Please share: We are #Moms4GMOs! Scientist and Advocate Moms to Celeb Moms: Weigh GMO Food with Facts Not Fear	Female	advocate	fear, garbage
cann4ing	BERNIE's ready to take on the GMO problem. Hillary's supported by evil MONSANTO!	Male	ready, supported	evil
_el_burro	I agree! Heard from a few people in Mexico expressing concern over this change in gmo ruling	None		concern

4. VOTING MECHANISM

When I finish the previous twitter processes, I went through the data again and found a problem. Some tweets posted by the same user who was classified as both female and male by Genderizer, it was consider as a false positive. This chapter describes how it happened, how I deal with this problem, and finally it shows us the statistical result.

4.1. Genderizer Mechanism

Let me show some more detail information about Genderizer. We can separate the Genderizer as following two parts:

4.1.1. Looking at a Username

It has a database store bunches of name and corresponding genders, which includes five different situations.

- "M": this name is definitely used for only male.

- "F": this name is definitely used for only female.

- "?M": mostly this name is a male's name.

- "?F": mostly this name is a female's name

- ?: It doesn't infer any gender for sure, when looking at this name

Table 4. Example of Username and Corresponding Gender

Username	Gender
Adam	M
Damiana	F
Dana	?F
Hette	?M
Manu	?

Figure 4 shows more detail procedure about how the Genderizer tries to determine a gender, which involves username detect and Naïve Bayesian Classification.
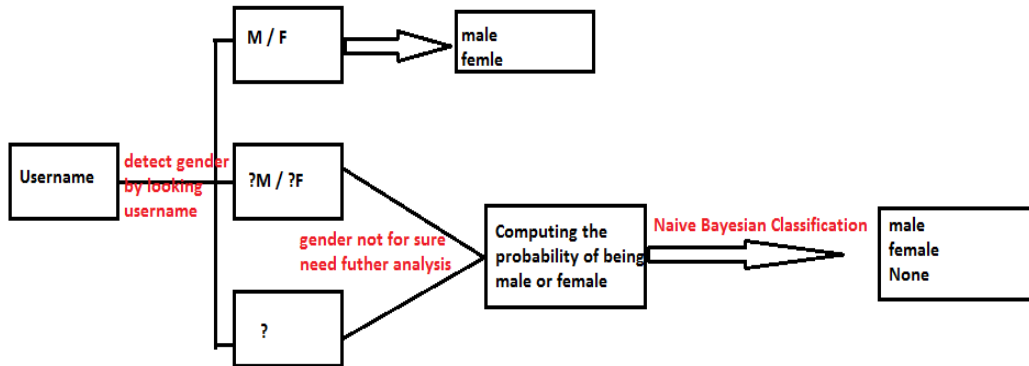

Figure 4. Detail Genderizer Processing

If the inference is either M or F then the tweeter is classified as a male or female respectively and it won't do any further tweet analysis. However if the username doesn't infer any gender for sure, then further tweet analysis need to be done.

The following figure shows the process about genderizer determines a gender by a username is definitely used for only one gender.
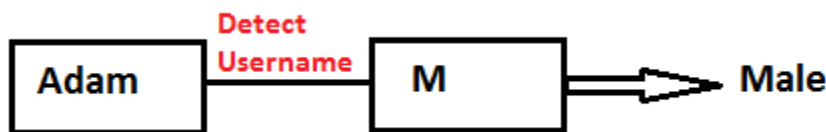

Figure 5. Example of Username is Definitely Used for Only Male

4.1.2. Detecting the Corresponding Tweet

When we detect the gender as "?M", "?F", or "?". Then tweet's text is used to find the gender. There are two different cases to compute the probability of being male or female mining the tweets.

4.1.2.1. Case 1

At the previous step a gender of username is "?M" or "?F", then only if the probability of the gender being a male or a female by mining the tweet larger than 0.6 then we classify the tweeter as either a male or a female.

The process is about how genderizer determine a gender from a username, which is identified as "?M" or "?F" as the following figure.
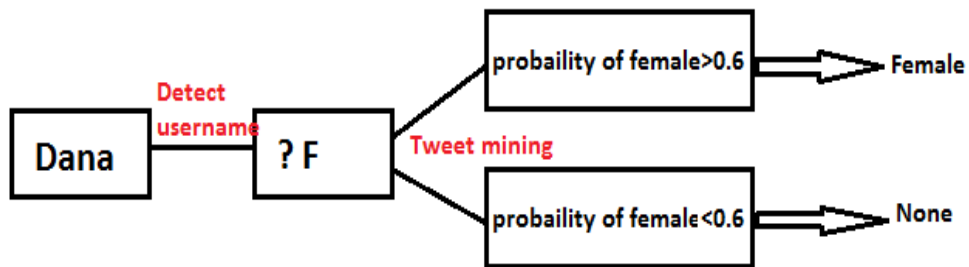

Figure 6. Example of Username is Mostly Used for Female

4.1.2.2. Case 2

If the gender of username is detected as "?", we should calculate the difference between the probabilities of male and female. Only if the difference is larger than 0.3 then we classify the tweeter as the gender with higher probability.

The process is about how genderizer determine a gender from a username, which is identified as "?" as the following figure.
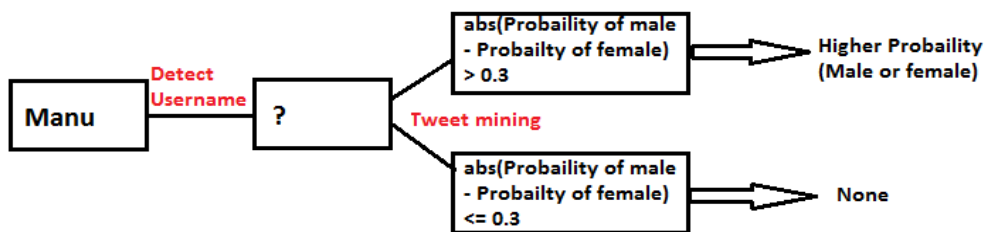

Figure 7. Example of Username doesn't Infer any Gender for Sure

14

4.2. Voting Scheme

So far, we have successfully finished the gender classification step. However, we found some tweeters post several tweets and some of the users were classified as both male and female. So I implemented the majority-voting scheme in order to deal with this problem. As the previous table, the user 8extremes was classified as female, male, and None. Firstly, I tried to look at if the user re-tweets any tweets. The third tweet shows us "RT" so it means this tweet was not originally posted by 8extremes so I discard the whole record. After that I tried to count how many males and females we have then chosen the majority gender as this user's real gender. For this example, the gender for 8extremes would be female.

Table 5. Example of the Same Tweeter was Classified as Different Genders

UserName	Tweet	Genderizer
8extremes	US: Add Scotland to the list of non-GMO countries! …	female
8extremes	best reasons 2 boycott gmo in us products but not eu -traitor brand…	male
8extremes	~~RT @swhotmess: 9. GMO will slowly KILL you.~~	None
8extremes	World Expert > Glyphosate definitely genotoxic & damages human DNA …	female

The following figure shows the process about how the voting scheme works based on majority mechanism.
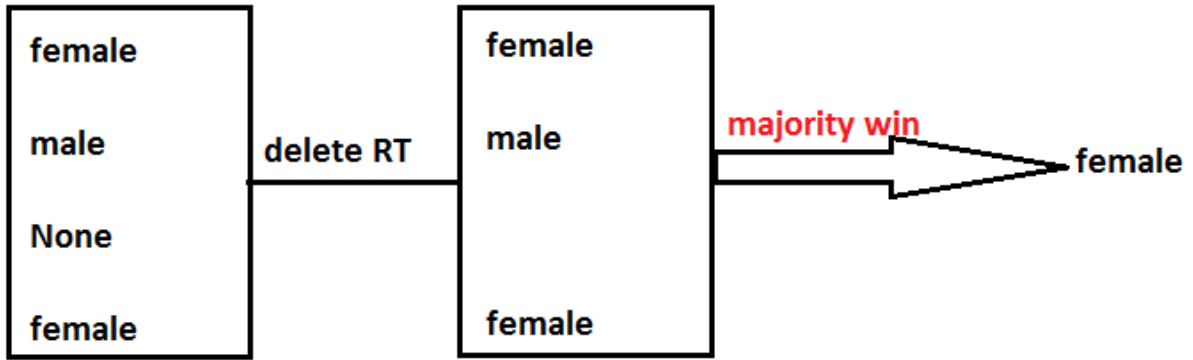
Figure 8. Process of Voting Scheme

5. STATISTICAL ANALYSIS

5.1. Data of Identify Gender

After all the previous steps, we got 2254 male tweets and 2589 female tweets in total 7557 tweets, since we only focus on the gender differences so we decided to discard all None tweets. Moreover, we also discard those tweets that did not consist of any positive word or negative word. Figure 10 illustrates that we used total 884 unique Twitter users for analysis, 39.59% males and the rest are females.
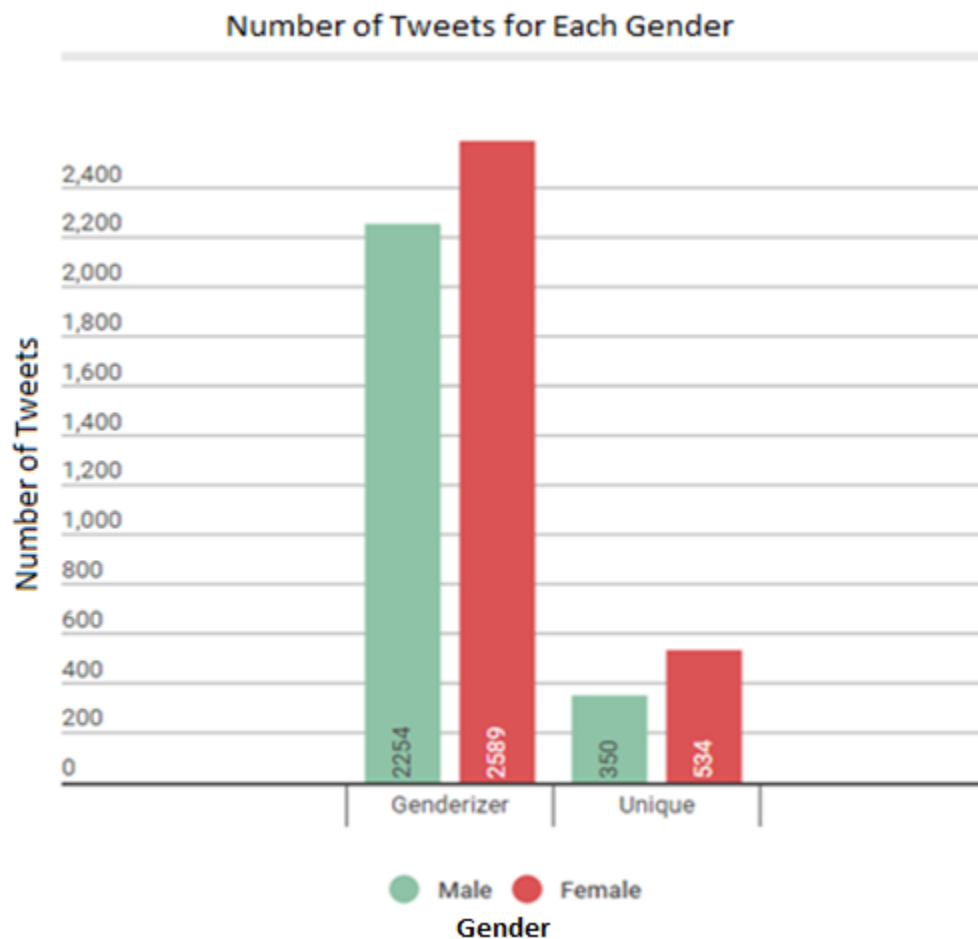
Figure 9. Numbers of Tweets for Each Gender

5.2. Positive and Negative Words Statistical

Following charts shows the most frequently pro-GMO and anti-GMO words extracted from the tweets for both male and female.

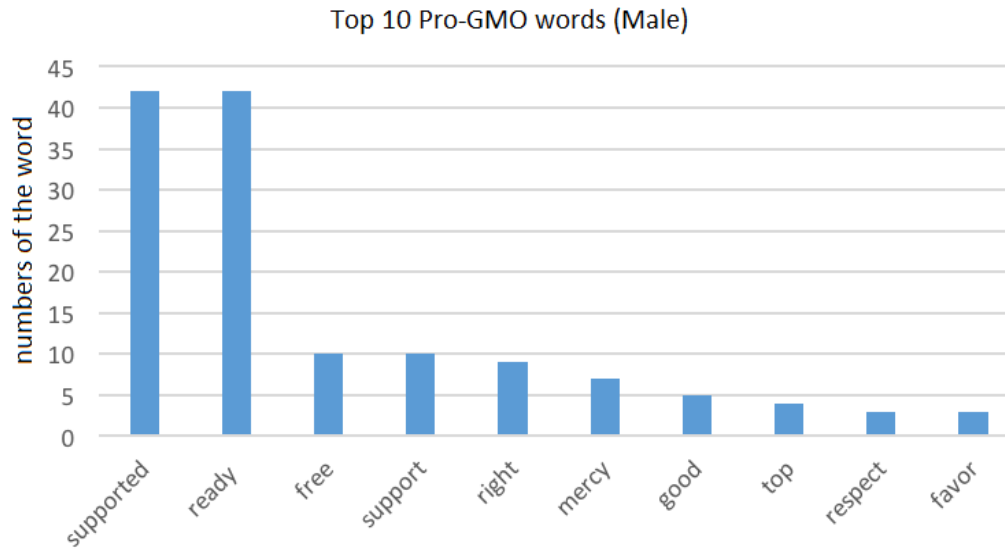Figure 10 shows the most appeared positive words in all male tweets.

Figure 10. Top 10 Male's Pro-GMO Words

Figure 11 shows the most appeared negative words in all male tweets.
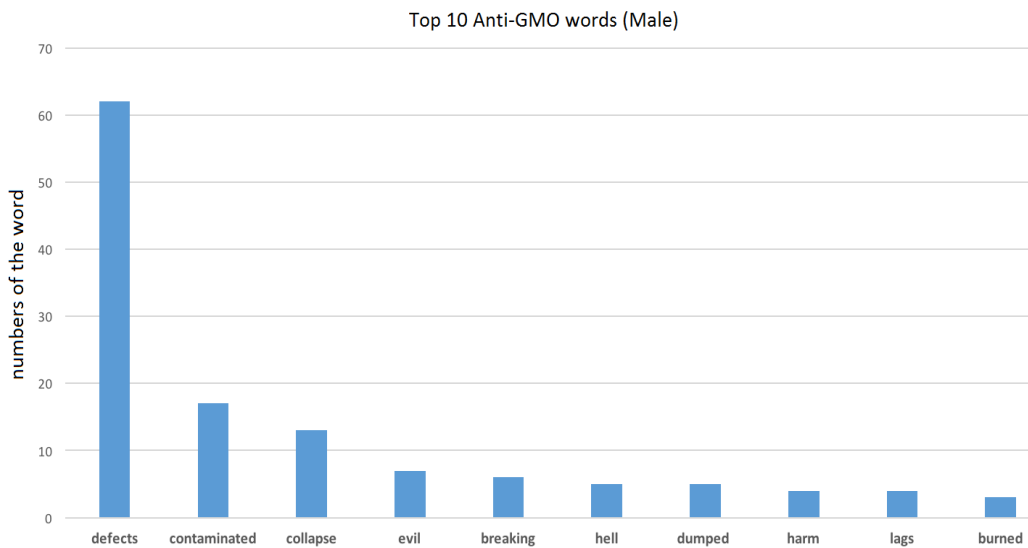

Figure 11. Top 11 Male's Anti-GMO Words

Figure 12 shows the most frequently Positive words in all female tweets.
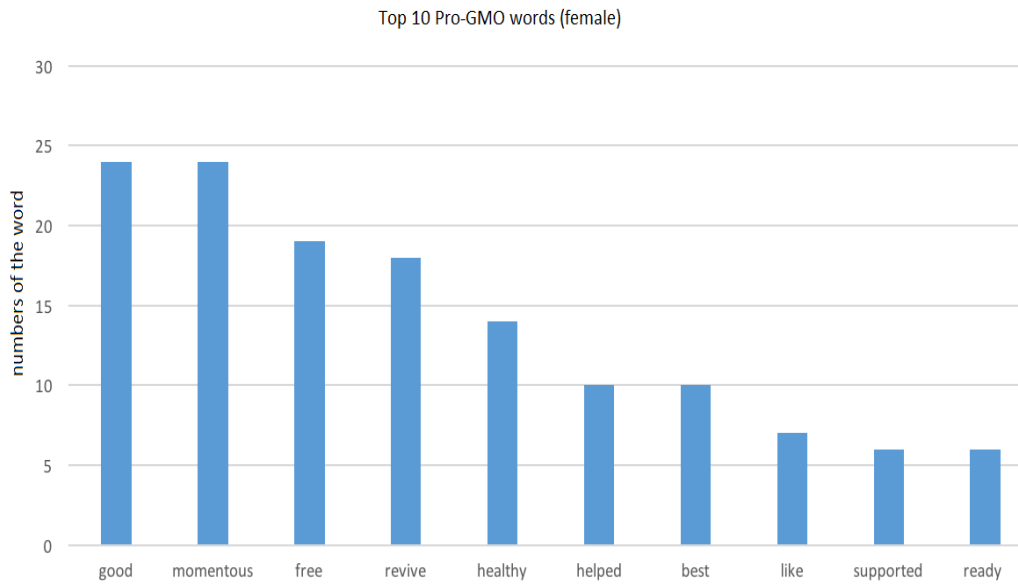

Figure 12. Top 10 Female's Pro-GMO Words

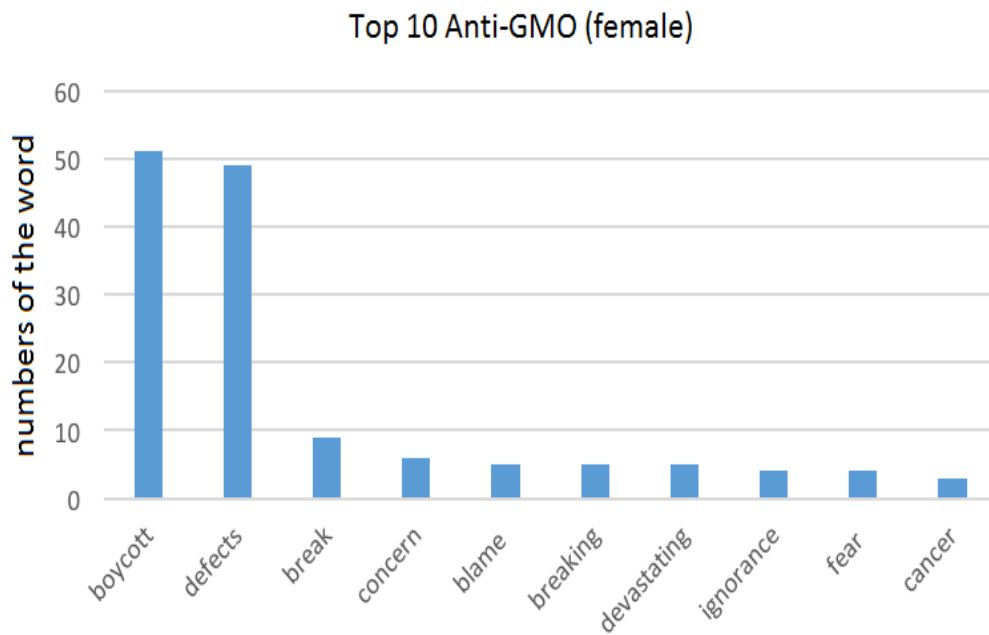Figure 13 shows the most frequently Negative words in all female tweets.

Figure 13. Top 10 Female's Anti-GMO Words

5.3. Statistical Test Result

Based on all previous experimental steps, the result has demonstrated that males have more negative opinions about GMOs when compared to all the male tweets (39.59%) and female tweets (60.41%) extracted from Twitter. The following figures illustrate the association between gender and nature of the GMO tweet.
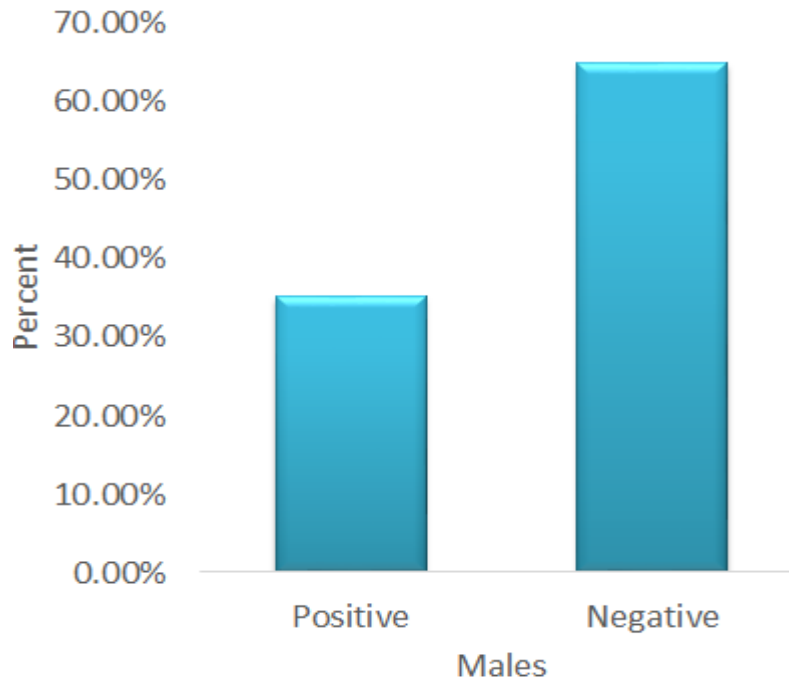
Figure 14. Percentage of Positive and Negative Opinions for GMOs in All Male Tweets

Figure 15 shows us the comparison of the percentages of negative emotions for GMOs in male and female tweets.
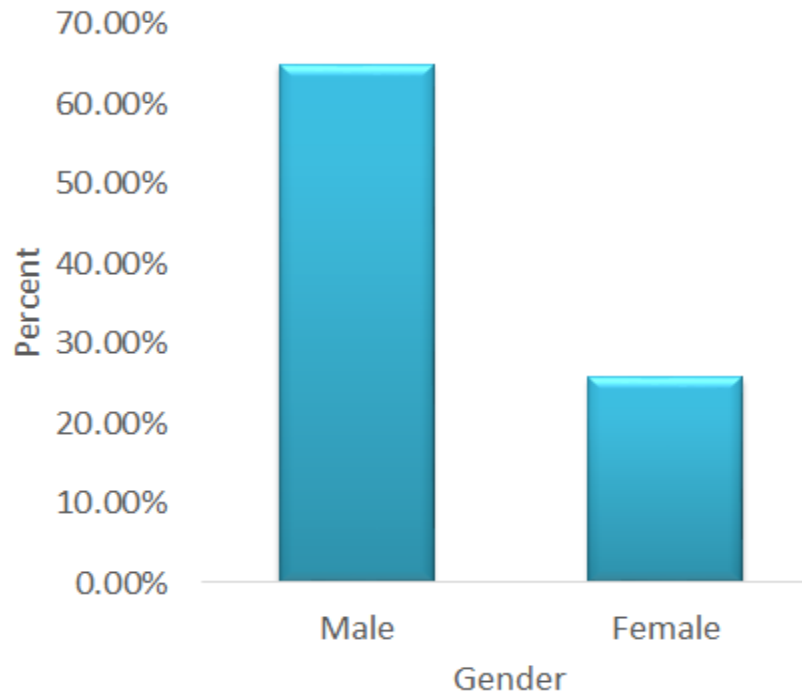

Figure 15. Percentage of Negative Opinions for GMOs in Both Male and Female Tweets

At last, we ran a chi-square test and according to the result of the test (p<0.0001), then we can say that there is a strong association relationship between gender and the nature of the tweets.

6. CONCLUSION AND FUTURE WORK

In this chapter, it gives an overall review and also describes how we can extend the study in the future.

6.1. Conclusion

Overall, for the work of this paper, our aim is to understand people's opinions concerning GMOs and whether there are differences by gender. So we designed the experimental and followed all the steps.

Firstly, we got all Twitter API keys: API key, API secret, Access token and Access token secret. After that, we can connect to Twitter Streaming API and start to capture Twitter feeds. Next step is split username and tweets then parse them to the Genderizer to classify genders. In order to improve the accuracy of Genderizer, I implemented the voting scheme and it helped a lot to reduce false positives. Meanwhile, I also extracted both positive and negative words from each tweet by matching the context in each single tweet with the word list. The last step was running statistical tests to determine the association between gender and the perceptions expressed in each single tweet.

Finally, the experimental result has demonstrated that males have more negative opinions about GMOs when compared in all the male tweets and female tweets extracted from Twitter. Meantime, the result of the chi-square test also validates this conclusion and since the value of p is much less than 0.001 so it infers that there is a strong association between gender and the nature of the tweets.

6.2. Future Work

6.2.1. GMO Twitter Analysis

In the future, I think we can extend this Twitter analysis to conduct relevant sentiment analysis. Based on this study, we can announce that gender differences play a role in people's perceptions of GMOs, but there are still some false positives. Because there are still some limitations here for example: sometimes two negative words could make a very positive sentence, but right now we can't figure out this problem. So if we can run some sentiment analysis then it could help us to improve the accuracy and reduce those false positives.

When we extracted tweets through the Streaming API, we also got some users location by latitude and longitude so I think we can do some more analysis based on locations as well. For example we've started to do some analyses based on people's perceptions of GMOs in the US and out of the US.

6.2.2. Twitter Analyses on Other Aspects

For some commercial or advertisement companies, they can do twitter analyses to understanding people's purchasing attitudes for their products. They can get some feedbacks and improve the quality of their products. Also, as I mentioned at the very beginning, social network makes the world smaller so it's a very good source for us to discover what is happening in the world and what people's opinions about those things. For example, investment corporations can use twitter to find out investment confidence for investors. Or even government can use those data to improve some policies to make public more satisfied with their lives.

REFERENCES

[1] High-level expert forum – How to feed the world in 2050: 'Global agriculture towards 2050', at World Summit on Food Security (Oct 12-13 2009).

[2] Pranav Dass, Md Chowdhury, Damian Lampl, Kendall E.Nygard. 2015 "Risk Perceptions for Genetically Modified Organisms: An Empirical Investigation", at the 34th IASTED International Conference.

[3] *The Streaming API overview* Retrieved from https://dev.twitter.com/streaming/overview (Accessed: 2016-3-14)

[4] *Genderizer in Python* Retrieved at https://pypi.python.org/pypi/genderizer (Accessed: 2016-3-16).

[5] Dr. Jacqueline McLaughlin. 1996. "Independent Research Project: chi-square test".

[6] *Hypothesis Testing (P-value approach)* Retrieved at https://onlinecourses.science.psu.edu/statprogram/node/138 (Accessed: 2016-3-16).

[7] Montserrat Costa-Font, Jose´ M. Gil, and W. Bruce Traill. 2007. "Consumer acceptance, valuation of and attitudes towards genetically modified food: Review and implications for food policy" Food Policy, Volume 33 (Issue 2) Pages 99-111

[8] A. Agarwal and J. Sabharwal. 2012. End-to-end sentiment analysis of Twitter data. In proceedings of the workshop on Information Extraction and Entity Analytics on Social Media Data, at the 24th International Conference on Computational Linguistics (IEEASMD-COLING '12).

[9] Pranav Dass, Yang Lu, Md. Chowdhury, Damian Lampl, Janani Kamalanathan, Kendall E. Nygard, 2016, Gender Differences in Perceptions of Genetically Modified Foods. 31st

International Conference on Computers and Their ApplicationsCATA 2016, Las Vegas, Nevada, USA April 4-6, 2016

[10] Genderize.io *Determine the gender of a first name* Retrieved from https://store.genderize.io/documentation (Accessed: 2016-3-16)

[11] Dan Nguyen. 2015. *Creating the Genderizer; a Lookup Table to Guess Gender from Names* Retrieved from http://2015.padjo.org/tutorials/babynames-and-college-salaries/creating-genderizer-table-with-joins/

[12] I. Rish. 2001 An empirical study of the naïve Bayes classifier, In proceedings of IJCAI-01 workshop on empirical methods in AI, Sicily, Italy, pp. 41-46.

[13] M. Hu and B. Liu, Mining and summarizing customer reviews, in proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04), Seattle, Washington, USA, 2004.