Project for C7084 Big Data and Decision Making - Case Studies

Georgina Wager

12/04/2021

This project is part of the assessment for the module C7084 Big Data and Decision Making - Case Studies, that is ran at Harper Adams University College by Ed Harris for the MSc Course: Data Science for Global Agriculture, Food and Environment.

The task of this project is to use data science techniques to obtain and analyse data.

Github Name: georginaanna Repository Name: C7084-Assesment-Big-Data Github Link direct to the repository:https://github.com/georginaanna/C7084-Assesment-Big-Data (https://github.com/georginaanna/C7084-Assesment-Big-Data)

Project: Understanding the opinions surrounding veganism

Latest reports by the vegan society have claimed that there were around 600,000 vegans in Great Britain in 2019, an increase in comparison to previous years (The Vegan Society, 2021). There has also been a report that documented 172 registered crimes against vegans because of their beliefs between 2015 and 2020, demonstrating the importance of understanding societies' thoughts and opinions to combat and reduce discrimination. At the beginning of 2020, to provide vegan equality, a tribunal ruled that ethical veganism is a philosophical belief, which means it is protected by law under the Equality Act 2010 (Nachiappan, 2020).

An article by Brown (2019) highlighted some issues on how the opinions of vegans is impacting the mental health of farmers documenting the multiple occurrences of online abuse with many landowners and butchers having concerns about their safety and privacy due to vegan activists uploading an interactive map online showing their addresses and contact details.

In addition to this, Harper Adams University, a robust agricultural university, was advertising "Veganuary" in January 2021, this received a negative response from the farming community and revoked an apology from the university, as reported by both Henderson (2021) and FarmingUK (2021). The argument here is should an apology have been issued primarily when open debate should be allowed, especially with Harper Adams being considered the future of food and farming.

Social media is the fuel to the fire and is often a haven for fake news; in the author's own beliefs and experiences, many UK farmers feel their industry has been misrepresented through social media. However, this project aims to better understand the conversations surrounding the vegan topic to identify the negative and positive terms used and understand the feelings surrounding veganism. Hopefully, this will enable us to understand people's thoughts towards vegans and prevent the hatred from both parties that have been witnessed over recent years.

Twitter began in 2006 and is a known form of social networking. Twitter allows users to share limited character messages on their user profiles that are shared publicly; these messages are known as tweets. Twitter has been chosen as the data source for this analysis due to its ability to allow for widespread communication between people known or unknown to share a topic of interest in everyday lives (Boyd, 2011). The purpose of Twitter is to distribute information; this information does not have to be factual (Parmelee and Bichard, 2011). As of September 2019, Twitter comprises 330 million monthly active users that post 500 million tweets per day.

The aim of this project is gain a better understanding of the opinions and thoughts surrounding veganism through using sentiment analysis. Sentiment analysis is the use of natural language processing (NLP) and text analysis to identify words within text and quantify them with regards to there sentiment.Some words can be

labeled simply through identifying if that word is negative or positive while in other cases the words can be quantified from -5 to +5. The main aim of sentiment analysis is to determine the emotional tone behind the text to gain an understanding of opinions and emotions.

Objectives:

1.Classify and identify words from tweets into positive or negative.

2.Identify the feelings associated with veganism.

3.Identify twitter parameters (retweet count, follower count, etc) that may influence a single tweet's sentiment value.

Method for obtaining the Twitter data

Firstly, the packages "rtweet", "httr" and "tidyverse" were installed. The next step was to apply for a Twitter API; a developer twitter account was needed to carry out this project. Using google, this URL pulled up the page that provides a further link to applying for a developer account https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api (https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api). This involved answering some basic questions so Twitter could ensure that what the data was going to be used for was within their data usage policies.These policies can all be found here (https://developer.twitter.com/en/developer-terms (https://developer.twitter.com/en/developer-terms)). The data questions include providing information on the core use/intent of the Twitter APIs, if there will be any interacting with tweets through liking and retweeting, the intended analyse of the Tweets, Twitter users, or content and if the Twitter content will be displayed. The "standard" Twitter API was the one that was applied for; this is specifically for one project and is free. To find historical tweets, a "tweet lookup" was carried out.

Once the account was approved, the app could be made by going to the Twitter developer dashboard, under the "app/create app" section. The name and the description fields were filled out, and the website URL is the Twitter account. This provides multiple codes that were inputted into the r script. Ones that were noted down and kept safe are the app name, the consumer key, the consumer secret, the access token, and the access secret. Once these were defined in R, a twitter token was created.

Next, the data was collected by identifying the words that were required in each tweet. Only tweets that have either of these words will be pulled into R. These words were identified as "vegan" or "vegans" but not both. Using the "search_tweets" function, the data was captured by defining the words required ("vegan" or "vegans") the number of rows (n=10000), not to include any retweets (include_rts = FALSE) and only tweets in the English language (lang="en"). Only 10000 rows were pulled. To demonstrate the amount of available data, this pulled tweets made on the 1st and 2nd of April 2021 with Tweets only containing the words "vegan" or "vegans". It is clear how big this data source is and supporting the need for a subsample.

Following this, the Twitter data was turned into a data frame. At this point, it contained 10000 rows and 90 columns. To reduce the dimensionality of the data set, only six columns were selected, the "display_text_width" that represented the width of the text, the "text" column that contained the information that was tweeted in text form, the "favorite_count" column, that counted the number of times that particular tweet had been favourited and the "retweet_count" column that counted the number of times that specific tweet had been retweeted. This was followed two columns that provided information on the persons account who made the tweet by the "followers_count" and the "friends_count". The data frame was then saved as an excel through the "write.xlsx" function. The detailed r script on obtaining the Twitter data is available on the authors GitHub: https://github.com/georginaanna/C7084-Assesment-Big-Data/blob/main/Dummy%20Getting%20the%20data.R (https://github.com/georginaanna/C7084-Assesment-Big-Data/blob/main/Dummy%20Getting%20the%20data.R)

Method for analysing the Twitter data

Firstly, the data was read into R via the “read.csv” package, and three sentiment data sets were obtained through the “textdata” package. The text data sets were “bing”,”afinn” and “nrc”. Each of these data sets is considered a lexicon. A lexicon is a dictionary of words that computes a word’s sentiment by analysing the “semantic orientation” of that word in a text. The “bing” dataset is a general-purpose English sentiment lexicon that categorises words in a binary fashion, either positive or negative. The data set contains the sentiment for 6,786 words and consists of two columns, one column contains the word, and the other includes the sentiment (positive or negative) (Bing and Hu, 2004). The “afinn” data set has 2,477 words; each word is given a number between -5 and 5, where minus five means that the word is very negative and five means that the word is very positive (Finn Årup Nielsen, 2011). The “nrc” data set contains 13,901 rows of words, and each word is assigned one of the eight primary emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (positive and negative) (Mohammad and Turney, 2010).

The tweets were then prepared using tokenisation through the function “unnest_tokens”, this means splitting each text word in the “text” column into small pieces called tokens; each word now has its own row, which has created a new data set called “unnest_tweets”. The “unnest_tweets” data set contains 231,555 rows of data (a row for all words that were in the text of every tweet).The “inner_join” function was used on the unnest_tweets dataset to join the “word” column in the bing data set. This then counted the number of times each word was used throughout the entire tweet data. This was used to create a word cloud grouping the positive and negative words, and the size of the word represented how often that word was found in the text (the bigger the word the higher presence of that word throughout the tweet data). To gain a better insight into the emotions surrounding the vegan topic, the “nrc” data set was used to group the number of words by feelings; this was then plotted by the frequency of each expression on multiple bar plots using the ggplot package.

Next a sparklyr connection was set up. Sparklyr is an R interface for Apache Spark. Apache Spark is an open-source, distributed processing system for analysing big datasets. It utilises in-memory caching and optimised query execution for quick analysis. Sparklyr package was installed from cran and a local version of Spark was installed for development purposes using the function “spark_install”. Next, a connection was made to a local instance of Spark via the “spark_connect” function. A graph was created to quantify the feelings associated with the tweet data. This was done like the previous graphs by using the “inner_join” functions to merge the “nrc” dataset, and the number of words with that specific sentiment were counted. Then the afinn data set was used as this provides 11 different categories for each word, providing more information on how positive or how negative each word is. This information was plotted on a bar chart.

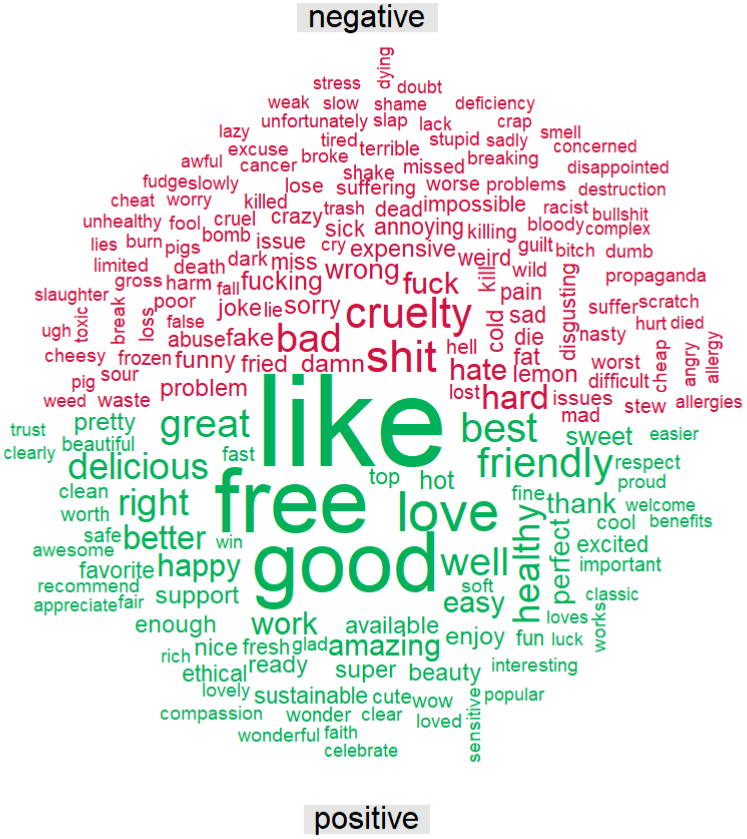
The previous analysis provided an insight into the sentiments used in tweets, how frequent these words were used and how positive or negative they were using the afinn and bing sentiment lexicons. The analysis also identified the words that are associated with certain feelings and the most common feelings based on the using the nrc dataset. The next step was to see if certain tweet parameters (retweet count, follower count, etc) had any influence on sentiment value. This was done by a multiple linear regression analysis fitted with a dependent variable and multiple independent variables (Chamber, 1992).

Two linear models were produced, one that looked at the effect that followers count, and friends count had on sentiment value and another that looked at text width, favourite count, and retweet count.The linear models were put together using the “ml_linear_regression” function, for each model the “value” of the sentiment is the response variable, this is followed by a tilde(~), and then the remaining independent variables. In addition, the data was split into a train and test dataset, this was carried out to be able to test the model accuracy on data using the “ml_regression_evaluator” to be able to obtain a Root Mean Squared (RMSE) (Witten and Frank 2005). The next step was to disconnect the spark connection.

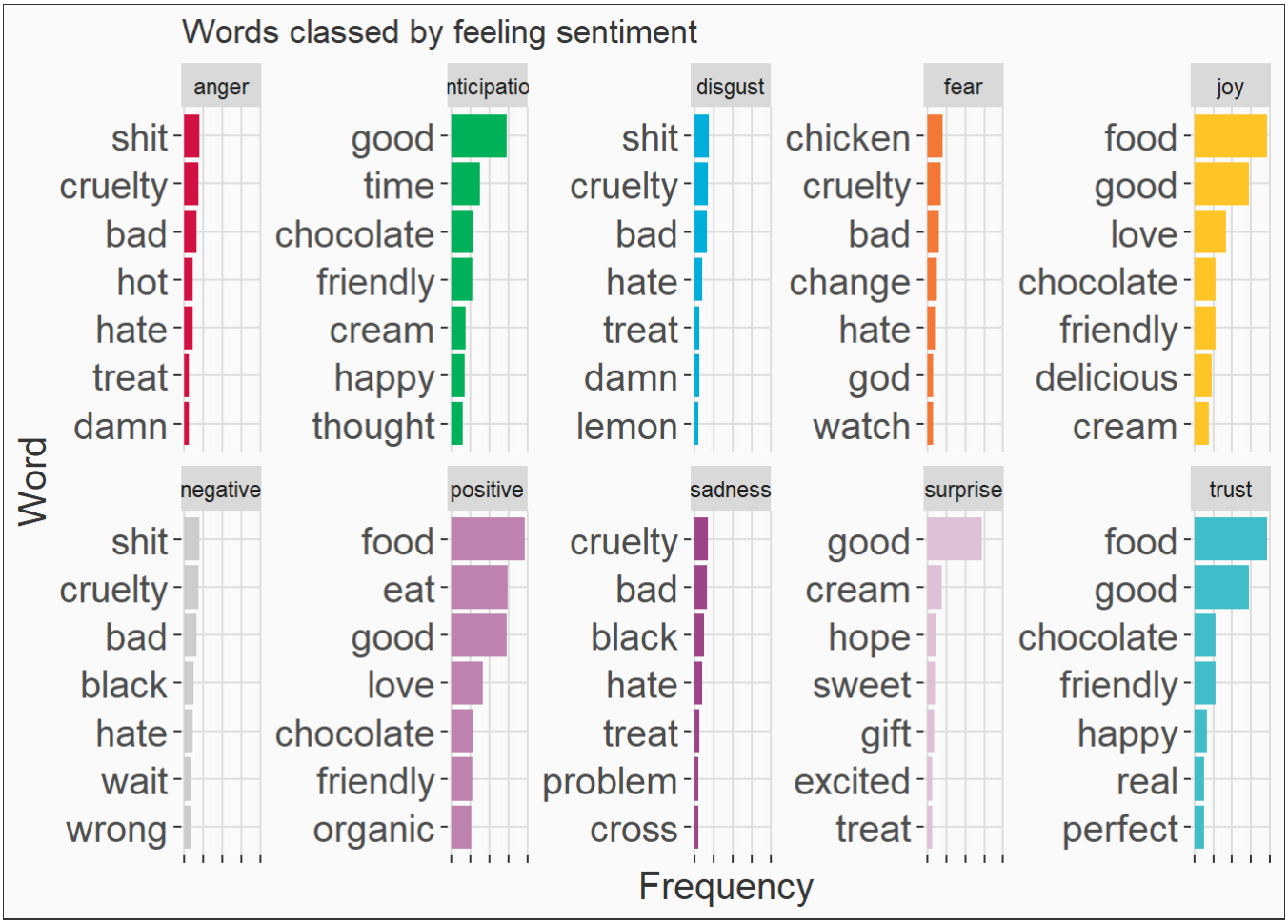
Results

The word cloud of negative and positive words showed that the most common negative words were “cruelty”, “shit”, “hard” and “bad”. The least common negative words used were “abuse”, “expensive” and “deficiency”.The most common positive words throughout the tweet data were “like”, “good”, “free” and “love”.

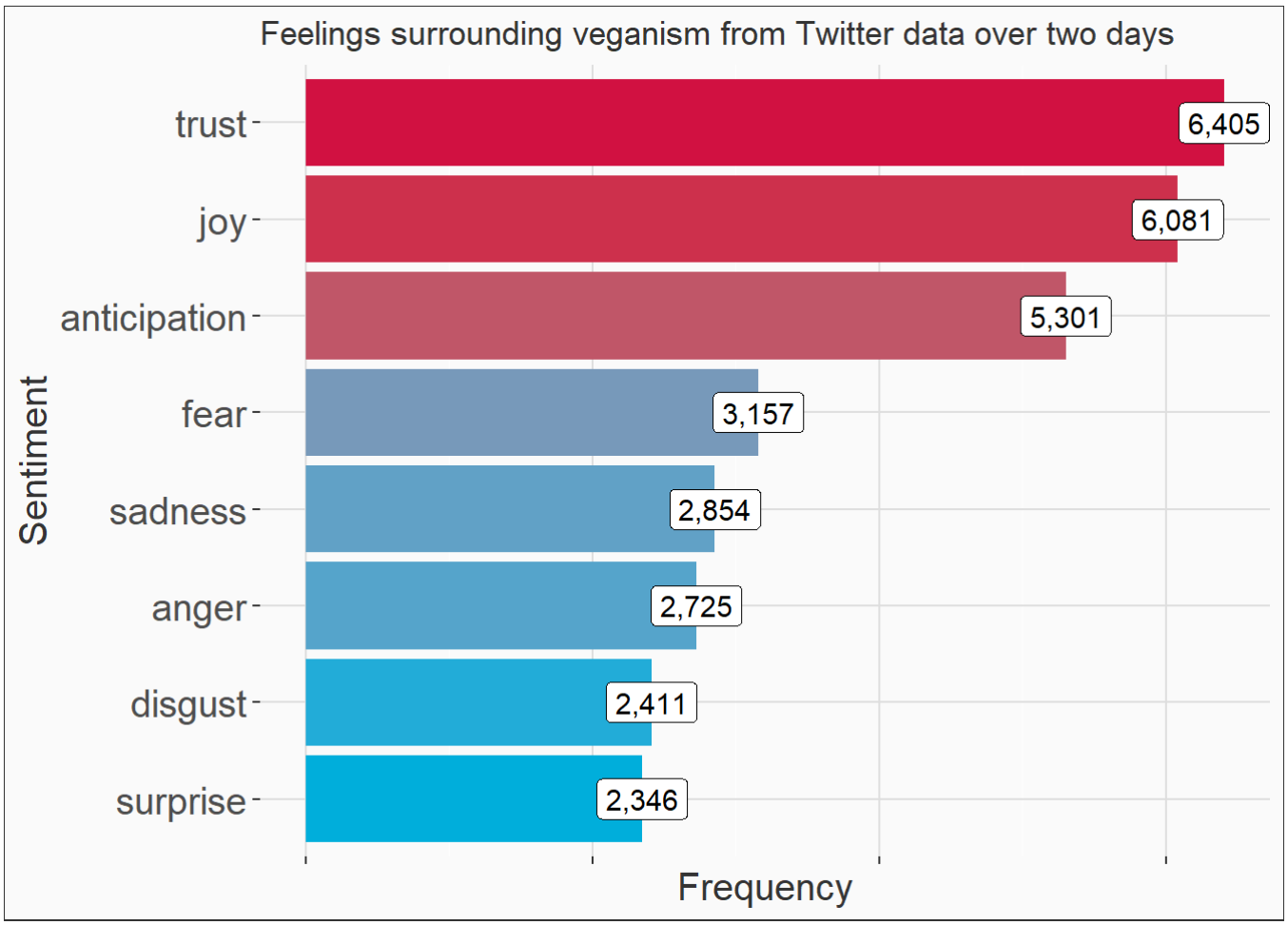
The least common positive words used were “respect”, “loves”, and “celebrate”.



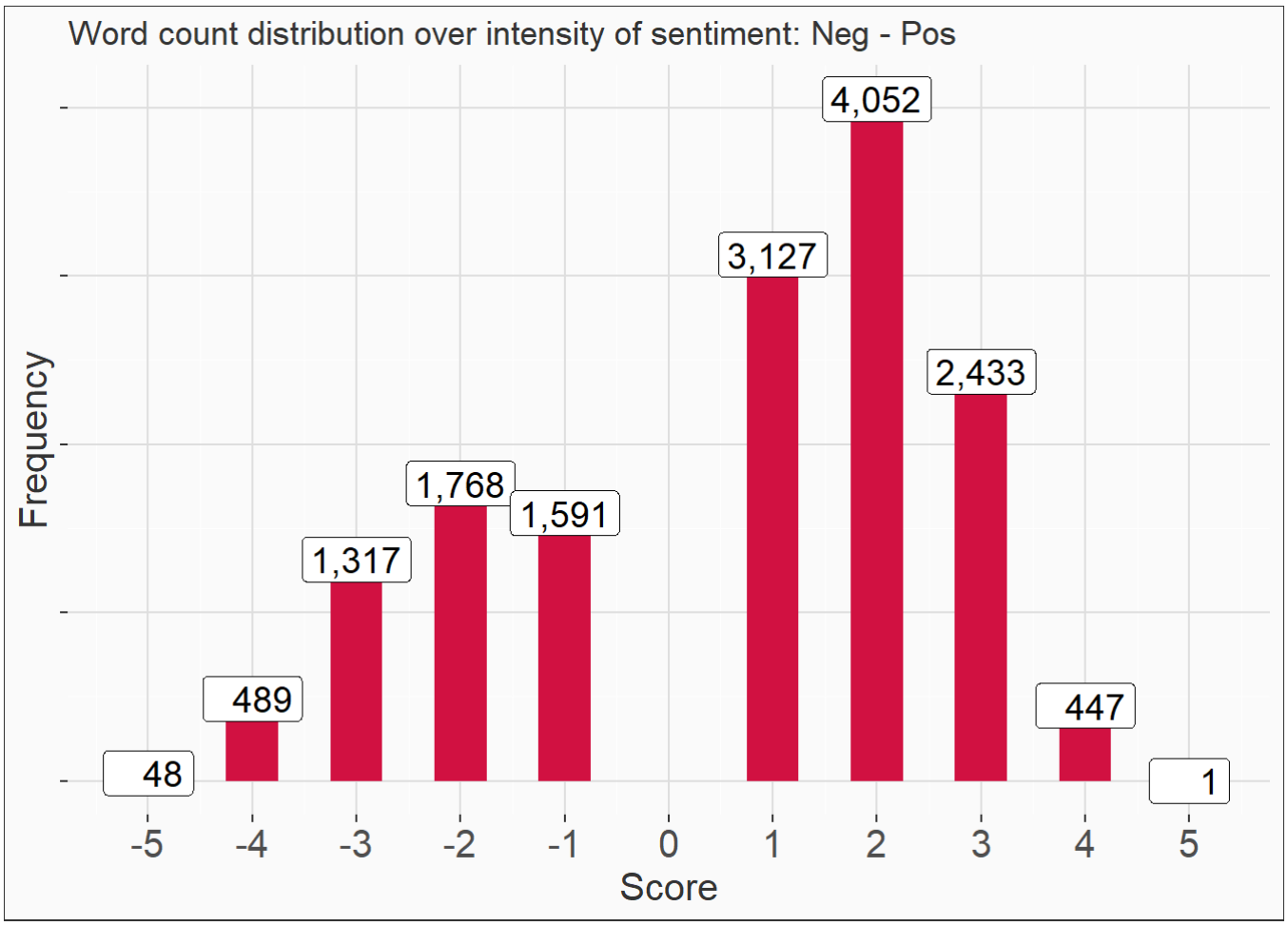
Overall, for the “Words classed by feeling sentiment” graph the highest proportion of words fell in positive, joy and trust sentiment. The smaller proportion of words were classed in the sentiments “disgust”, “negative”, “fear” and “anger”. When looking at the postive words for joy they seem to be explaining the food, insinuating that a vegan diet brings many people joy.



Overall, from the tweet data, 6,405 of the words used represented the trust emotion, 6,081 words represented joy, and 5,301 represented anticipation. In total, 31,280 words were labeled as having an emotion, and 40% of these words represented trust and joy which are considered quite positive feelings. However, 17% of the words were recognised through anticipation which is an emotion that can be perceived as either negative or positive. This is similar for the 7.5% of words that represented surprise. The remaining 35.5% were certainly negative emotions (fear, sadness anger, disgust). The balance of negative and positive emotions is fairy even regarding veganism.



Sentiment score “2” had the greatest word count, meaning words were mostly positive, and the bars are higher for positive sentiments. However, there are more words that are classed as extremely negative compared to extremely positive. This demonstrates the number of extreme words that are used when opinions are made regarding the vegan topic.



The results from the first linear regression model shows that the user metric “friends_count” has a significant effect with a p value of 0.00075, (p value<0.01) on the value of sentiment. For every one percent increase in “friend_count”, there is an associated 1.5 percent increase in sentiment score, however, “followers_count” was insignificant in relation to sentiment value (p-value>0.01).

The results from the second linear regression model show that retweet_count, favourite_ count and display_text_width is all less than p>0.01 and are therefore insignificant and have no impact on sentiment value.

Discussion

Despite the word cloud being a good visualisation, there is some limits with interpreting the data. For example, the word “free” has been classed as a positive word, but there is no context, as if this were a vegan talking about feeling free when taking a vegan diet or is it regarding freeing animals which would be a negative feeling from a vegan. Therefore, this adds a level of unreliability when using sentiment analysis to classify words. With the “Words classed by feeling sentiment” graph the highest proportion of words fell in positive, joy and trust sentiment, in comparison to more negative feelings, this demonstrated that during the time period of the tweets there was more good than bad words used surrounding the vegan topic. This is different to what would initially be expected. The following graph showed a more even balance between negative (anger, disgust, fear and sadness) and positive emotions (trust and joy). The affin data set concluded that the marjotiry of the words were positive compared to negative with a sentiment value of two, however, more extreme negative words were used surrounding the vegan topic compared to extremely positive, but it was still less than the amount of overall positive words at the positive end of the scale.

Overall the words were classified and identified with the most common negative words being “cruelty”, “shit”, “hard” and “bad”. The least common negative words used were “abuse”, “expensive” and “deficiency”.The most common positive words throughout the tweet data were “like”, “good”, “free” and “love”. The least common positive words used were “respect”, “loves”, and “celebrate”. The key feelings associated with

veganism were trust and joy and the least common from this data set were disgust followed by surprise. From the linear regression models it demonstrated that the number of friends people had on twitter influenced the sentiment value. This could possibly be marketing account for a company that promotes vegan foods, this wouldn't be surprising as when looking at the words from the second graph (Words classed by feeling sentiment) the top severn words for joy were "food", "good", "love", "chocolate", "friendly", "delicious" and "cream" which could all be associated with food. However ,it could easily by a genuine person who has lots of friends on there twitter profile that loves promoting vegan foods.

To gain a better understand with this analysis a few improvements could be made to the model, such as:

• Looking at more than one word to gain a better context such as the word "free" does this simply mean someone is feeling free, or does it mean animals should be free from slaughter or does it mean foods free from animal products. As this word was used a lot as seen in the word and classed as positive this could have affected the overall results.

• Removing accounts that are marketing vegan foods to remove any large opinion biased that could distort the data.

• Gaining a sentiment value for each user instead of the overall text used to understand more individual opinions.

• Pull together tweets from a larger sample size, instead of two days.

• The dates for this data were the 1st and 2nd of April, which was two days before Easter Sunday, with Easter Sunday being the usual day to eat lamb this could have increased the use of the "cruelty" in individuals tweets.

Despite the above points, for the time frame used the actual emotions surrounding the vegan topic are fairly positive, which is often not what it seems from social media postings as outlined at the beginning of this project.The negative words are more from what a vegan would say about why they are vegan with the words "cruelty" and "slaughter", and not what a non vegan would likely say, concluding that from this data there is reason to believe that the majority of the opinions surrounding the vegan topic that are both positive and negative are from vegans. Understanding the thoughts and opinions of others and identifying negative words used in tweets could help social media outlets to have a better control over there users, especially when certain social media postings surrounding the vegan topic has had a direct impact on individuals mental health and well being.Hopefully further work can be carried out to help prevention to protect those who choose different life choices whether that be eating animal product or not.

References

Boyd, D., Golder, S. and Lotan, G., 2010. January. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In 2010 43rd Hawaii international conference on system sciences. pp. 1-10. IEEE.

Brown. 2019. Online abuse and farm protests: the vegans impacting on farmers' mental health. [Online]. The Sustainable Food Trust. Available from: https://sustainablefoodtrust.org/articles/online-abuse-and-farm-protests-the-vegans-impacting-on-farmers-mental-health/ (https://sustainablefoodtrust.org/articles/online-abuse-and-farm-protests-the-vegans-impacting-on-farmers-mental-health/) [Accessed 25 March 2021].

Chambers, J. M. 1992. Linear models. Chapter 4 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

FarmingUK. 2021. Harper Adams Students' Union backtracks after vegan post. [Online]. Available from: https://www.farminguk.com/news/harper-adams-students-union-backtracks-after-vegan-post_57404.html (https://www.farminguk.com/news/harper-adams-students-union-backtracks-after-vegan-post_57404.html) [Accessed 8 April 2021].

Henderson, E. 2021. Harper Adams Students' Union blasted for Veganuary post. [Online]. MA Agriculture Ltd, a Mark Allen Group company. Available from: https://www.fwi.co.uk/news/harper-adams-students-union-blasted-for-veganuary-post (https://www.fwi.co.uk/news/harper-adams-students-union-blasted-for-veganuary-post) [Accessed 8 April 2021].

Hu, M. and Bing, L. 2004. Mining and summarising customer reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004), Seattle, Washington, USA, Aug 22-25, 2004.

Mohammad, S. and Turney, P. 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, June 2010, LA, California.

Nachiappan, A. 2020. Experts get their teeth into idea of vegan hate crime. [Online]. The Times. Available from: https://www.thetimes.co.uk/article/experts-get-their-teeth-into-idea-of-vegan-hate-crime-65nsf6c02 (https://www.thetimes.co.uk/article/experts-get-their-teeth-into-idea-of-vegan-hate-crime-65nsf6c02) [Accessed 8 April 2021].

Nielsen, F. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718 in CEUR Workshop Proceedings 93-98.

Parmelee, John H, and Shannon L Bichard. 2011. Politics and the Twitter Revolution: How Tweets Influence the Relationship Between Political Leaders and the Public. Lexington Books.

The Vegan Society. 2021. Worldwide. [Online]. The Vegan Society. Available from: https://www.vegansociety.com/news/media/statistics/worldwide (https://www.vegansociety.com/news/media/statistics/worldwide) [Accessed 6th April 2021].

Witten I. and Frank E., 2005. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA