Classification of Mixed Martial Arts' Tweets using Logistic Regression

Yu Jen Chang
Department of Computer Science
University of Glasgow
Glasgow, G12 8QQ
2582255c@student.gla.ac.uk

June 2022

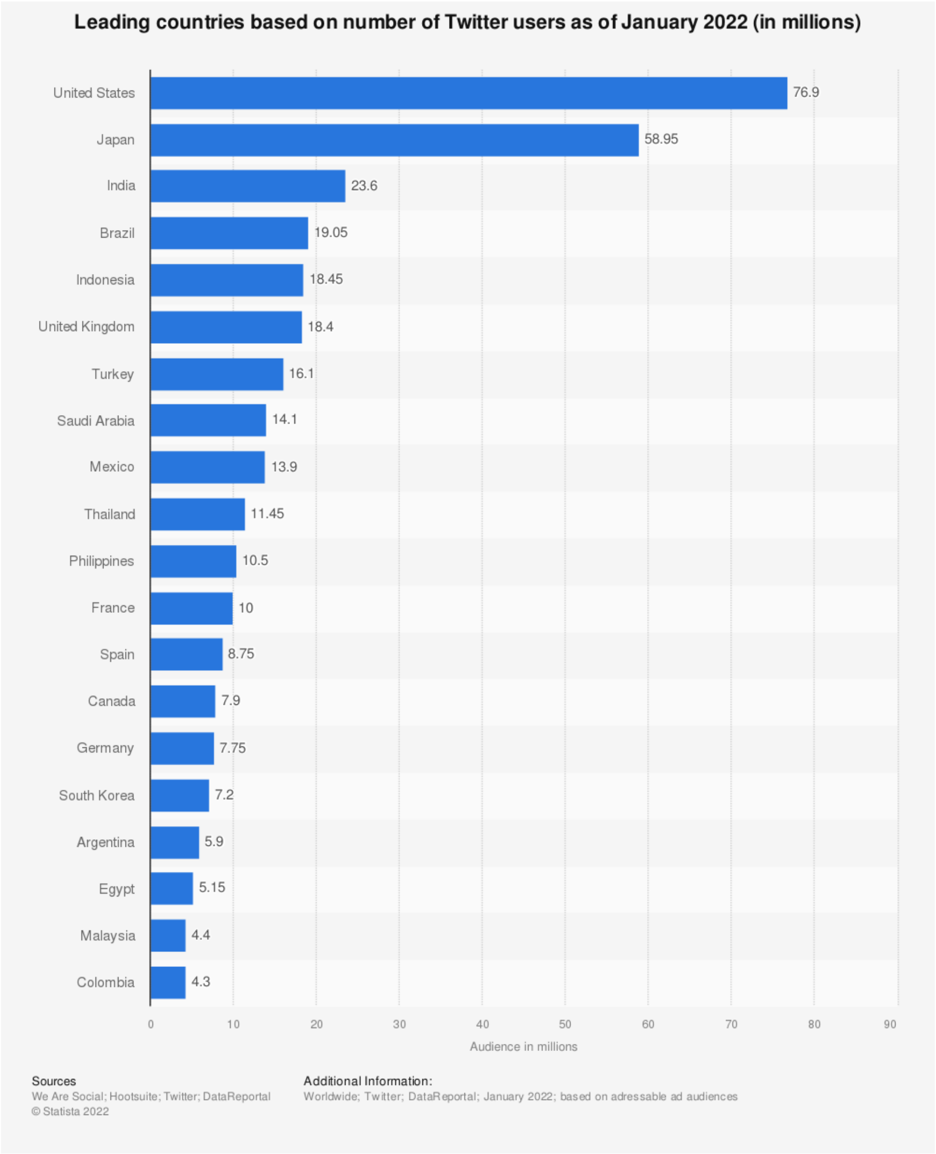
Abstract

In an age when free speech is highly valued, it has become mainstream to express one's opinion openly on social media. People express their opinions on the internet, including about restaurants, politics, or sporting competitions. However, some of the comments are filled with hate speech and toxic language. Sentiment analysis has emerged in recent years as a way of distinguishing between positive and negative content; it not only helps people to be aware of toxic language but also helps companies to target their advertising more precisely to generate better revenue. This dissertation focuses on the use of sentiment analysis to analyse tweets related to the mixed martial arts (MMA) games, including tweets from MMA fans, athletes, and anyone talking about the MMA game, using various machine learning and sentiment analysis methods such as SVC, LSTM, Naive Bayes, word embedding, etc. In this project, binary classifiers classify toxic language in MMA tweets as positive and negative, in addition, multi-classifiers classify negative tweets into different categories.

1 Introduction

Social media has become a source of information for modern society, allowing people to post their opinion in real time, on topics such as sports, politics, etc. Twitter is one of the major social media platforms, where users can post tweets of up to 280 characters in length. Due to the ease of use and rapid spread of Twitter, Twitter will have over 300 million users worldwide by 2022, with the largest number of users in North America[1].

[1] https://www.statista.com/statistics/303681/twitter-users-worldwide/



Sources
We Are Social; Hootsuite; Twitter; DataReportal
© Statista 2022

Additional Information:
Worldwide; Twitter; DataReportal; January 2022; based on adressable ad audiences

Figure 1: Histogram of countries based on number of Twitter users as of January 2022

The popularity and public nature of Twitter (less than 10% of Twitter accounts are private[2]) make it an important tool for studying people's behaviour and attitudes. One area of research that has attracted a great deal of attention in the last few years is the sentiment classification of Twitter[15]. Sentiment categorisation and analysis allows one to learn about people's attitudes towards specific topics on Twitter. There have been a number of works that have classified sentiment on Twitter using standard sentiment classification techniques, the most common of which are varia-

[2] https://techcrunch.com/2009/10/05/twitter-data-analysis-an-investors-perspective-2/

tions on the n-gram and bag of words, eg., Kouloumpis et al. (2011)[11], Ren et al. (2016)[17] and Denecke (2008)[3]. There have also been attempts to use more advanced syntactic features, e.g., user information[20].

Most sentiment analysis of tweets focuses on reviews of movies, restaurants, or products. However, To the best of our knowledge, no one has used sentiment analysis of tweets related to MMA. This project focuses on tweets from MMA. As there have been many MMA fights throughout history, the focus of this project is one of the best-selling fights in UFC history which is UFC229[3], collecting and analysing tweets from fighters and hashtag tweets related to the fight.

MMA, also known as cage fighting or ultimate fighting, is a full-contact combat sport based on boxing, grappling, and ground fighting. MMA was first marketed as a competition to find the most effective martial arts for actual unarmed combat, pitting competitors from various fighting styles against one another in matches with few rules.

The Ultimate Fighting Championship (UFC) is an American MMA promotion company founded in 1993, also the biggest organization in the world, in twelve weight divisions (eight men and four women[4]). Until 2022, UFC has held over 500 events around the globe. Figure 2 shows a 400% growth in UFC twitter followers between 2015 and 2020.

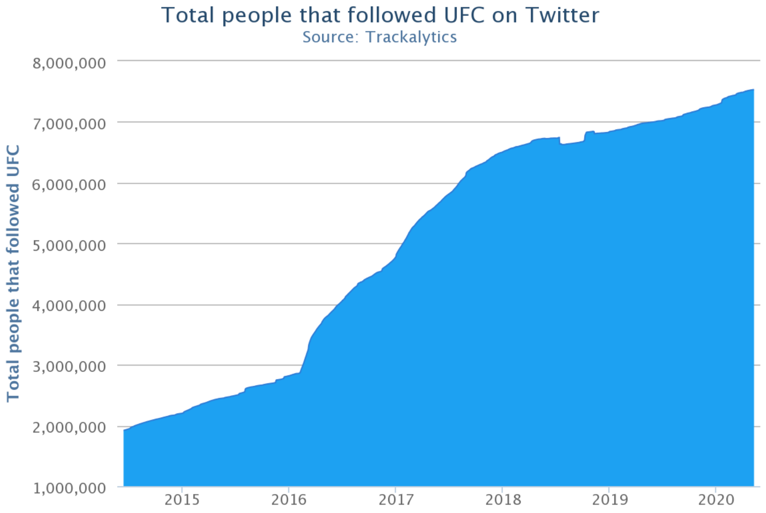


Figure 2: Follower Growth Number

MMA fighters, similar to wrestlers, need to establish symbols for their characters[1]. Fighters create conflict by speaking on social media, at press conferences or in front of the screen, which helps build the fullness of their characters to create buzz and engage audiences. One of the most famous fight in recent years was Khabib Nurmagomedov verse Conor McGregor in October 2018 for UFC229, a fight that sold 2,400,000 Pay-Per-View (PPV) for a total of at least $180 million[5].

Abusing opponents on social media has become a common way for fighters to create conversation

[3] https://sports.betmgm.com/en/blog/mma/biggest-pay-per-view-events-ufc-bm05/

[4] https://www.ufc.com/

[5] https://sports.betmgm.com/en/blog/mma/biggest-pay-per-view-events-ufc-bm05/

and conflict, This project focuses on sentiment analysis of MMA tweets data, observe tweets contain mostly obscene words and what types of toxic comments are targeted.

2 Related Work

Sentiment analysis and classification of text has been well researched and implemented in various topics and domains, and has led to new and different challenges. For example, Sida Wang et al. used NB-SVM to provide a baseline for sentiment analysis and topic classification[19]. Go et al's. sentiment analysis of Twitter data has led to effective results in the processing of emoticons in the data [6]. The use of neural networks to classify hate speech by Björn et al. shows remarkable results and implementation. [4]. Not only did the methods used by Björn et al. and Ramish et al. perform well in their respective studies, but Nhan et al.'s study also synthesised excellent findings that the use of RNN and word embedding methods performed well across many different data sets[8][2]. However, Aditya et al. also showed that Logistic Regression gave the best classification results when using Logistic Regression, Naive Bayes, and Support Vector Machines as classifiers respectively[5].
In this project I tried three different types of models as baselines: NB-SVM, Logistic Regression with TF-IDF, and LSTM with Word Embedding. Due to hardware and time constraints, the LR model was chosen as the baseline model for binary classification and multi-label classification.

2.1 Naive Bayes and Support Vector Machine (NB-SVM)

In text categorization and sentiment analysis research, NB-SVM models are frequently utilised as starting points for additional techniques. One study suggested a straightforward version of the model in which the SVM was constructed using the NB logarithmic ratio as the eigenvalue and shown that it was a powerful and reliable performer in each of the tasks given[19]. The primary model versions in the Sida Wang et al's. research are expressed as linear classifiers, with the following prediction for test case k is

$$y^{(k)} = sign(\mathrm{w}^T x^{(k)} + b)$$

For the SVM, $x^{(k)} = \hat{f}^{(k)}$, w, b, are obtained by minimizing

$$\mathrm{w}^T\mathrm{w} + C\sum_I max(0, 1 - y^{(i)}(\mathrm{w}^T\hat{f}^{(i)} + b))^2$$

Sida Wang et al's. result find the L2-regularized L2-loss SVM to work the best and L1-loss SVM to be less stable[19]. In MNB, $x^{(k)} = \hat{f}^{(k)}$, $\mathrm{w} = \mathrm{r}$ and $b = log(N_+/N_-)$, and N_+, N_- are the number of positive and negative training cases, and find that binarizing $f^{(k)}$is better. Sida Wang et al's. study also showed that NB did better than SVM in the sentiment task for fragments, but that SVM performed better for long texts. However, combining MNB and SVM with interpolation between the two for either sentiment or topic classification of fragments or long texts gave good performance. The report result model is

$$\mathrm{w'} = (1 - \beta)\bar{w} + \beta\mathrm{w}$$

This paper also give the result that bigram features in the bag of features have been relatively underestimated in previous studies.

2.2 Logistic Regression with TF-IDF Feature

Logistic Regression is the training of a classifier that makes a binary decision on the class of new input observations[10]. It is the sigmoid classifier that helps us to make this decision. A single input observation is first considered and a feature vector is used to represent it. The output of the classifier can be 1 (meaning the observation is a member of the class) or 0 (the observation is not a member of the class)[5].

By learning a vector of weights and a bias term from a training set, logistic regression completes this task. Each weight wi is a real number that corresponds to a particular input characteristicx_I. The weight wi, which can be either positive or negative (giving proof that the instance being classed belongs in the positive class), indicates how significant that input feature is to the classification decision (providing evidence that the instance being classified belongs in the negative class)[13].

$$Z = (\sum_{i=1}^{n} w_i x_i) + b$$

After learning the weights in training, the classifier multiplies each $x_i$by its weight $w_I$before adding the bias term $b$and making a judgement on a test instance. The weighted sum of the evidence for the class is expressed by the single number Z that results.

The Bag of Words technique, which is useful for text classification or for assisting a machine read words in numbers, is outperformed by the TF-IDF or Term Frequency(TF) - Inverse Dense Frequency(IDF) technique when it comes to understanding the meaning of sentences made out of words[16]. Term Frequency, is the relative frequency of term$t$within document$d$:

$$\text{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d}} f_{t',d}$$

Inverse document frequency is a measure of how much information the word provides:

$$\text{idf}(t,D) = log \frac{N}{|\{d : t \in d\}|}$$

The formula that is used to compute the TFIDF of term t present in document d is:

$$\text{tfidf}(d,t) = tf(t) * idf(d,t)$$

This method is based on extracting n-gram features from tweets and weighting them according to TFIDF, which has been achieved by reducing the effect of frequently occurring and less informative markers in the database on the training results. The performance of each algorithm was analysed by averaging the cross-validation scores for each combination of the feature parameters and then comparing the performance of the three algorithms.

2.3 LSTM with Word Embedding

Long short-term memory (LSTM) is a special type of Recurrent Neural Network (RNN) that is designed to solve the vanishing gradient and exploding gradient problems during the training of long sequences[18].

RNNs can link previous information to the current task, depending only on the most recent information. However, in semantic analysis, there are situations where more context is needed to make predictions, the entirely possible for the gap between the relevant information and the point where it is needed to become very large. Unfortunately, as that gap grows, RNNs become unable to learn to connect the information, and LSTM solves this problem of long term dependency. Figure 3 shows the basic LSTM architecture[7].

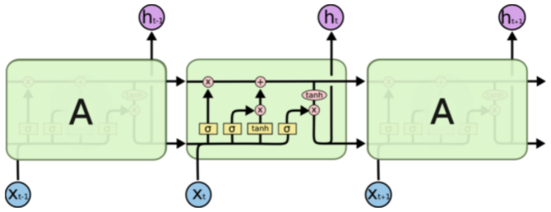


Figure 3: The repeating module in an LSTM contains four interacting layers

Word embedding is a vector representation that maps source data to another space by means of a function that converts words, sentences, etc. into a vector form. The processing of textual content is transformed into a vector operation in vector space to calculate the similarity in vector space[12].

The reason for converting text into spatial vectors is that in the bag-of-words model, there is a lack of understanding of the contextual relationships between words, whereas word embedding captures the context, semantic and syntactic similarities, and relationships with other words in the document[14]. Nhan et al. also show that word embedding methods perform better in comparison to TF-IDF using a neural network model[2].

Word2Vec is one of the most popular techniques for learning word embeddings using shallow neural networks[14]. It was developed by Tomas Mikolov at Google in 2013. In Word2Vec, there are two general models, Continuous Bag of Words (CBOW) and Skip-gram. As the Skip-gram model allows for better performance in semantic analysis, word vectors trained through the Word2Vec Skip-gram model are used as input to the classification phase.

Although the accuracy and practicality of this method has been mentioned in various papers, the training time is too long, ranging from 8 to 16 hours per session. With the limited time and hardware available, it is difficult to use it as a primary training method and is therefore not used in this project.

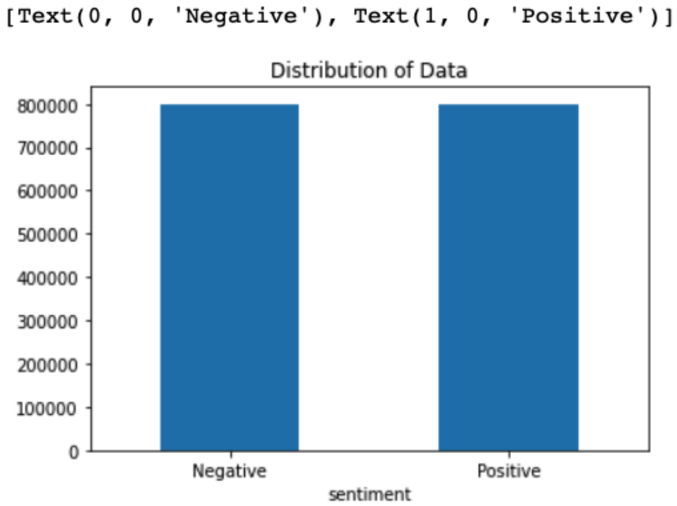
3 Data Collection and Description

The following three datasets were used in this projects: (1)the Sentiment140 dataset, (2)the Wikipedia Comments dataset and ()the MMA dataset, which I collected from Twitter.

- **Sentiment140 Dataset**

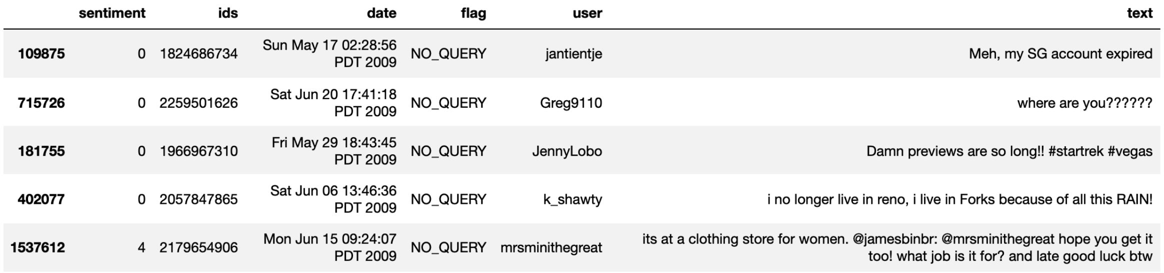
Stanford University provided Sentiment140[6]. There are 1.6 million tweets in it that mention brands or products. The polarity of the mood expressed by the author of each tweet was already indicated on the tweets (0 = negative, 4 = positive). It contains the following 6 fields: target (sentiment), ids, data, flag, user, and text.

This dataset is mainly used for training binary classification, so that the model can distinguish between negative and positive sentences. The figures below present the bar graphs of the Negative and Positive label in this data set and the sample of this data set.



```
[Text(0, 0, 'Negative'), Text(1, 0, 'Positive')]
```

Figure 4: Sentiment140 Labels Distribution



	sentiment	ids	date	flag	user	text
109875	0	1824686734	Sun May 17 02:28:56 PDT 2009	NO_QUERY	jantientje	Meh, my SG account expired
715726	0	2259501626	Sat Jun 20 17:41:18 PDT 2009	NO_QUERY	Greg9110	where are you??????
181755	0	1966967310	Fri May 29 18:43:45 PDT 2009	NO_QUERY	JennyLobo	Damn previews are so long!! #startrek #vegas
402077	0	2057847865	Sat Jun 06 13:46:36 PDT 2009	NO_QUERY	k_shawty	i no longer live in reno, i live in Forks because of all this RAIN!
1537612	4	2179654906	Mon Jun 15 09:24:07 PDT 2009	NO_QUERY	mrsminithegreat	its at a clothing store for women. @jamesbinbr: @mrsminithegreat hope you get it too! what job is it for? and late good luck btw

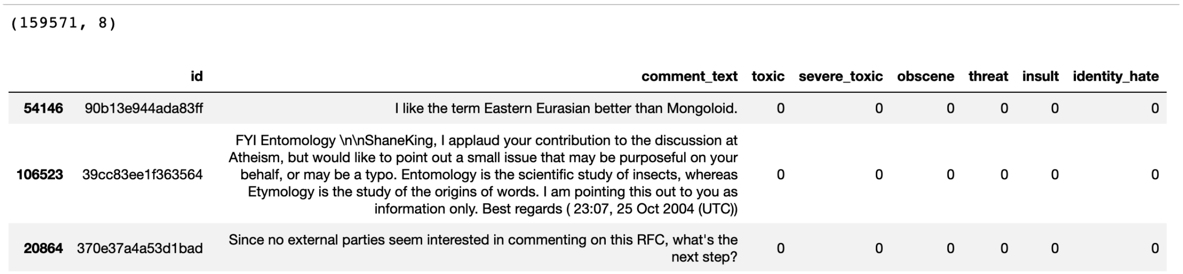
Figure 5: Random Sample of Sentiment140 Data Set

- **Wikipedia Comments**

The dataset is under CC0, with the underlying comment text being governed by Wikipedia's CC-SA-3.0. There are 159571 of Wikipedia comments data set which have been labeled by human raters for

[6] https://www.kaggle.com/datasets/kazanova/sentiment140

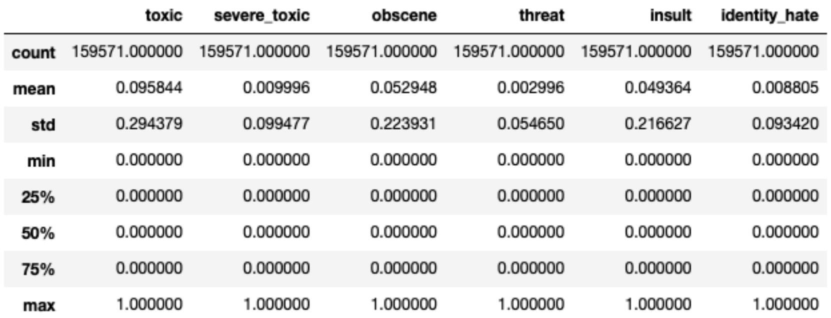
toxic behavior[**https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge**]. The types of toxicity are: toxic, severe toxic, obscene, threat, insult, and identity hate. This dataset was used to train a classification weight model for toxic languages, which was later tested using the MMA dataset as a test set. The figures below are a random sample and description of this data set.



```
(159571, 8)
```

	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
54146	90b13e944ada83ff	I like the term Eastern Eurasian better than Mongoloid.	0	0	0	0	0	0
106523	39cc83ee1f363564	FYI Entomology \n\nShaneKing, I applaud your contribution to the discussion at Atheism, but would like to point out a small issue that may be purposeful on your behalf, or may be a typo. Entomology is the scientific study of insects, whereas Etymology is the study of the origins of words. I am pointing this out to you as information only. Best regards (23:07, 25 Oct 2004 (UTC))	0	0	0	0	0	0
20864	370e37a4a53d1bad	Since no external parties seem interested in commenting on this RFC, what's the next step?	0	0	0	0	0	0

Figure 6: Random Sample of Wikipedia Comments Data Set

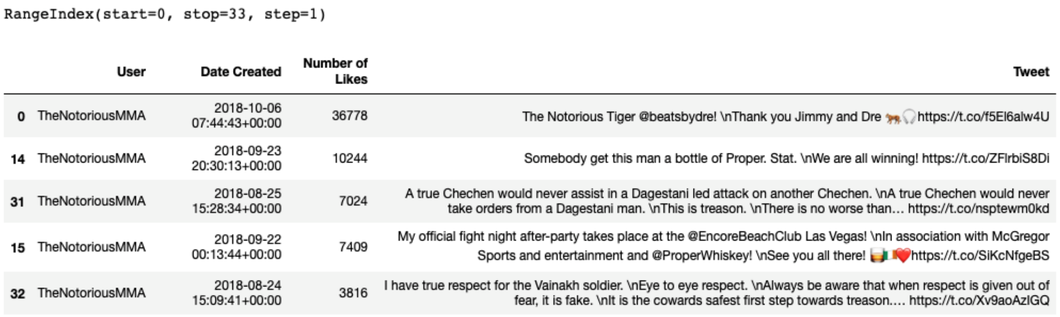


	toxic	severe_toxic	obscene	threat	insult	identity_hate
count	159571.000000	159571.000000	159571.000000	159571.000000	159571.000000	159571.000000
mean	0.095844	0.009996	0.052948	0.002996	0.049364	0.008805
std	0.294379	0.099477	0.223931	0.054650	0.216627	0.093420
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Figure 7: Wikipedia Comments Data Set Describe

- ### MMA Dataset

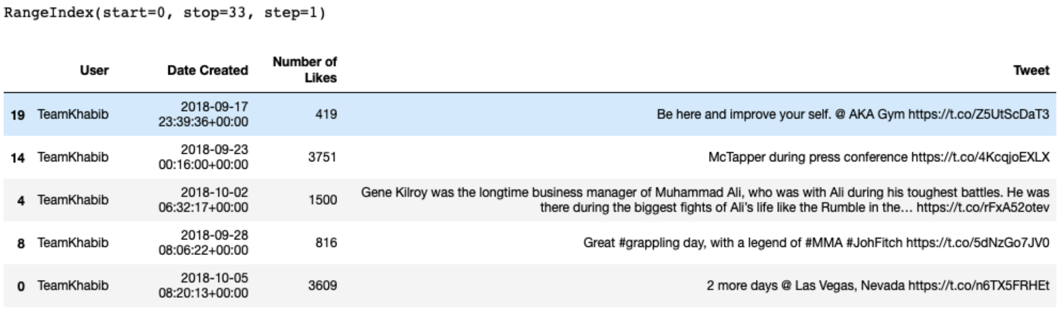
Since capturing the entire MMA data would be too large and randomly capturing it would lack thematicity, I chose one of the biggest selling major fights in UFC history, Khabib Nurmagomedov and Conor McGregor in October 2018 for UFC 229, as the main source of the data set. I started with a range of 45 days of tweets from 2018-08-23 to 2018-10-07, starting with when the official fight poster was first released and ending with the day after the fight. The user name and hashtag are used as keywords for crawling. I have captured the tweets of Khabib Nurmagomedov and Conor McGregor, the main contestants in this race, during this period. As Khabib's main language is Russian, this project focuses on analysing the sentiment categories in English, unfortunately I need to removal the Russian tweets. The chart below shows the tweets of the two contestants during this period, with Conor's total of 34 tweets and Khabib's total of 66 tweets contain the Russian and English tweets, and 34 only English tweets.



```
RangeIndex(start=0, stop=33, step=1)
```

	User	Date Created	Number of Likes	Tweet
0	TheNotoriousMMA	2018-10-06 07:44:43+00:00	36778	The Notorious Tiger @beatsbydre! \nThank you Jimmy and Dre 🐅🎧https://t.co/f5El6alw4U
14	TheNotoriousMMA	2018-09-23 20:30:13+00:00	10244	Somebody get this man a bottle of Proper. Stat. \nWe are all winning! https://t.co/ZFlrbiS8Di
31	TheNotoriousMMA	2018-08-25 15:28:34+00:00	7024	A true Chechen would never assist in a Dagestani led attack on another Chechen. \nA true Chechen would never take orders from a Dagestani man. \nThis is treason. \nThere is no worse than… https://t.co/nsptewm0kd
15	TheNotoriousMMA	2018-09-22 00:13:44+00:00	7409	My official fight night after-party takes place at the @EncoreBeachClub Las Vegas! \nIn association with McGregor Sports and entertainment and @ProperWhiskey! \nSee you all there! 🥃🇮🇪❤️https://t.co/SiKcNfgeBS
32	TheNotoriousMMA	2018-08-24 15:09:41+00:00	3816	I have true respect for the Vainakh soldier. \nEye to eye respect. \nAlways be aware that when respect is given out of fear, it is fake. \nIt is the cowards safest first step towards treason.… https://t.co/Xv9aoAzIGQ

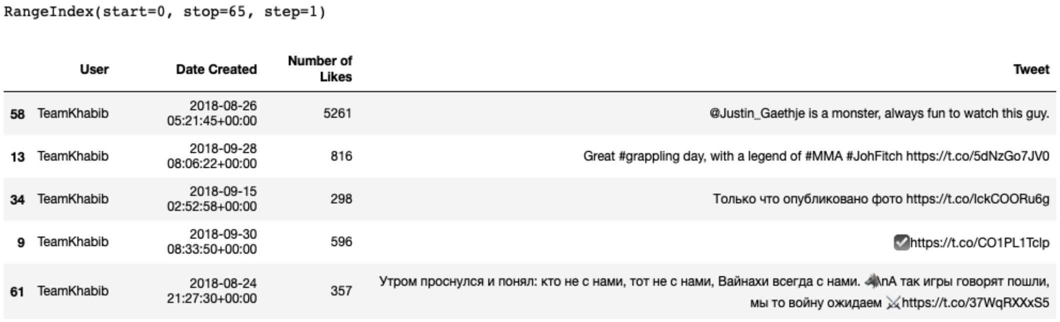
Figure 8: Conor McGregor Tweets



```
RangeIndex(start=0, stop=33, step=1)
```

	User	Date Created	Number of Likes	Tweet
19	TeamKhabib	2018-09-17 23:39:36+00:00	419	Be here and improve your self. @ AKA Gym https://t.co/Z5UtScDaT3
14	TeamKhabib	2018-09-23 00:16:00+00:00	3751	McTapper during press conference https://t.co/4KcqjoEXLX
4	TeamKhabib	2018-10-02 06:32:17+00:00	1500	Gene Kilroy was the longtime business manager of Muhammad Ali, who was with Ali during his toughest battles. He was there during the biggest fights of Ali's life like the Rumble in the… https://t.co/rFxA52otev
8	TeamKhabib	2018-09-28 08:06:22+00:00	816	Great #grappling day, with a legend of #MMA #JohFitch https://t.co/5dNzGo7JV0
0	TeamKhabib	2018-10-05 08:20:13+00:00	3609	2 more days @ Las Vegas, Nevada https://t.co/n6TX5FRHEt

Figure 9: Khabib Nurmagomedov Tweet in English



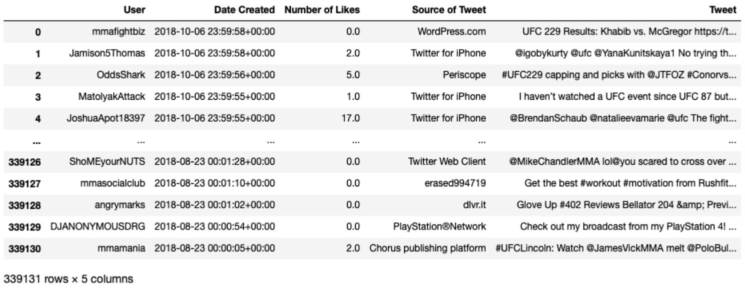
```
RangeIndex(start=0, stop=65, step=1)
```

	User	Date Created	Number of Likes	Tweet
58	TeamKhabib	2018-08-26 05:21:45+00:00	5261	@Justin_Gaethje is a monster, always fun to watch this guy.
13	TeamKhabib	2018-09-28 08:06:22+00:00	816	Great #grappling day, with a legend of #MMA #JohFitch https://t.co/5dNzGo7JV0
34	TeamKhabib	2018-09-15 02:52:58+00:00	298	Только что опубликовано фото https://t.co/IckCOORu6g
9	TeamKhabib	2018-09-30 08:33:50+00:00	596	✅https://t.co/CO1PL1Tclp
61	TeamKhabib	2018-08-24 21:27:30+00:00	357	Утром проснулся и понял: кто не с нами, тот не с нами, Вайнахи всегда с нами. 🐺\nA так игры говорят пошли, мы то войну ожидаем ⚔️https://t.co/37WqRXXxS5

Figure 10: Khabib Nurmagomedov Tweet

Hashtags are heavily used on twitter as a categorical tag and indicator for visitors to browse messages. I have chosen 4 hashtags as keywords to search for, namely ufc229, ufc, TheNotoriousMMA, and TeamKhabib.

The reason for this is that these are highly relevant tags to the event itself, with UFC being the organiser of the event and UFC229 being the number of the event, which is also officially used as the main name of the event. TheNotoriousMMA and TeamKhabib are the uesrname for the two fighter on Tiwtter, and many fans and other fighter will tag them on the Twitter at their own tweet for interaction. Using these four hashtags as keywords I gathered a total of 339,131 tweets.

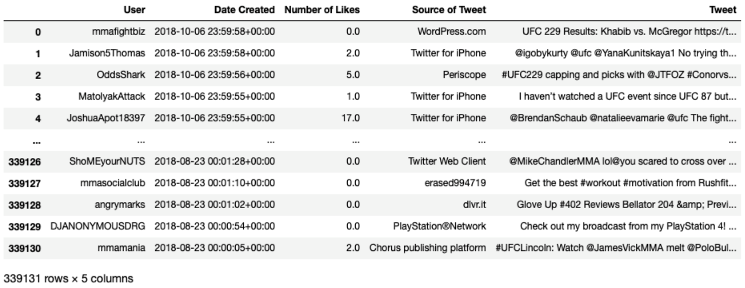


	User	Date Created	Number of Likes	Source of Tweet	Tweet
0	mmafightbiz	2018-10-06 23:59:58+00:00	0.0	WordPress.com	UFC 229 Results: Khabib vs. McGregor https://t...
1	Jamison5Thomas	2018-10-06 23:59:58+00:00	2.0	Twitter for iPhone	@igobykurty @ufc @YanaKunitskaya1 No trying th...
2	OddsShark	2018-10-06 23:59:56+00:00	5.0	Periscope	#UFC229 capping and picks with @JTFOZ #Conorvs...
3	MatolyakAttack	2018-10-06 23:59:55+00:00	1.0	Twitter for iPhone	I haven't watched a UFC event since UFC 87 but...
4	JoshuaApot18397	2018-10-06 23:59:55+00:00	17.0	Twitter for iPhone	@BrendanSchaub @natalieevamarie @ufc The fight...
...	...	...	...	...	...
339126	ShoMEyourNUTS	2018-08-23 00:01:28+00:00	0.0	Twitter Web Client	@MikeChandlerMMA lol@you scared to cross over ...
339127	mmasocialclub	2018-08-23 00:01:10+00:00	0.0	erased994719	Get the best #workout #motivation from Rushfit...
339128	angrymarks	2018-08-23 00:01:02+00:00	0.0	dlvr.it	Glove Up #402 Reviews Bellator 204 &amp; Previ...
339129	DJANONYMOUSDRG	2018-08-23 00:00:54+00:00	0.0	PlayStation®Network	Check out my broadcast from my PlayStation 4! ...
339130	mmamania	2018-08-23 00:00:05+00:00	2.0	Chorus publishing platform	#UFCLincoln: Watch @JamesVickMMA melt @PoloBul...

339131 rows × 5 columns

Figure 11: Tweets for 4 keywords

Finally, all search results are combined into a data set and remove the duplicate search results, showing a total of 338,882 tweets.



	User	Date Created	Number of Likes	Source of Tweet	Tweet
0	mmafightbiz	2018-10-06 23:59:58+00:00	0.0	WordPress.com	UFC 229 Results: Khabib vs. McGregor https://t...
1	Jamison5Thomas	2018-10-06 23:59:58+00:00	2.0	Twitter for iPhone	@igobykurty @ufc @YanaKunitskaya1 No trying th...
2	OddsShark	2018-10-06 23:59:56+00:00	5.0	Periscope	#UFC229 capping and picks with @JTFOZ #Conorvs...
3	MatolyakAttack	2018-10-06 23:59:55+00:00	1.0	Twitter for iPhone	I haven't watched a UFC event since UFC 87 but...
4	JoshuaApot18397	2018-10-06 23:59:55+00:00	17.0	Twitter for iPhone	@BrendanSchaub @natalieevamarie @ufc The fight...
...	...	...	...	...	...
339126	ShoMEyourNUTS	2018-08-23 00:01:28+00:00	0.0	Twitter Web Client	@MikeChandlerMMA lol@you scared to cross over ...
339127	mmasocialclub	2018-08-23 00:01:10+00:00	0.0	erased994719	Get the best #workout #motivation from Rushfit...
339128	angrymarks	2018-08-23 00:01:02+00:00	0.0	dlvr.it	Glove Up #402 Reviews Bellator 204 &amp; Previ...
339129	DJANONYMOUSDRG	2018-08-23 00:00:54+00:00	0.0	PlayStation®Network	Check out my broadcast from my PlayStation 4! ...
339130	mmamania	2018-08-23 00:00:05+00:00	2.0	Chorus publishing platform	#UFCLincoln: Watch @JamesVickMMA melt @PoloBul...

339131 rows × 5 columns

Figure 12: MMA Data Set

4 Resources and Pre-processing of data

Data pre-procession is an important step in an NLP project. The raw data is disorganised and contains a lot of unnecessary information that does not help us analyse the information and reduces the accuracy of the model[9].

A lot of tweets contain links to websites and user names that are not semantically meaningful and do not assist in semantic analysis. These low importance links and user names are replaced at this step. The website link replace to “URL” and the user name replace to “USER”.There is a lot of informal language in the tweet and a character is repeated a emphasis. For example: the positive word ‘good’ will be ”gooooood!”; the negative word ‘fuck’ will be ‘fuuuuuck’. The ‘o’ and ‘u’ are repeated more than twice in this situation, so any character repeated more than 2 times will be reduced to the remaining 2 times.

Emoticons in tweets are often ignored or deleted as special characters when they are processed, but in fact they also have a meaning that can help with sentiment analysis[6]. This project prepare the emoticon dictionary by labeling 170 emoticons listed on Wikipedia with their emotional state. For example: ”:)”, ”:-)”, ”:-D”, all of them represent smiling labels[7].

Finally, stopwords are removed using the stopwords package in Natural Language Toolkit (NLTK) and lemmatized using WordNetLemmatizer[8]. The reason for using lemmatization rather than stemming here is that lemmatization restores characters intact, whereas using stemming to cut characters causes both over stemming and under stemming errors [XX], which has its limitations. lemmatization is a better choice in this lemmatization is a better choice in this case.

5 Methodology and Result

5.1 Binary Classification

Firstly, we need to create a baseline binary classifier to classify the sentiment of tweets into positive and negative. In this baseline, I use three prominent machine learning algorithms as the tweet classifier: Bernoulli Naive Bayes, Linear Support Vector Classification, and Logistic Regression. The training data set is Sentiment 140 form Stanford University. Using the world cloud observed the clean data after pre-processing. According the figure, The substituted ’USER’ is heavily represented in both the positive and negative word clouds, with the negative word clouds having no overtly negative vocabulary.

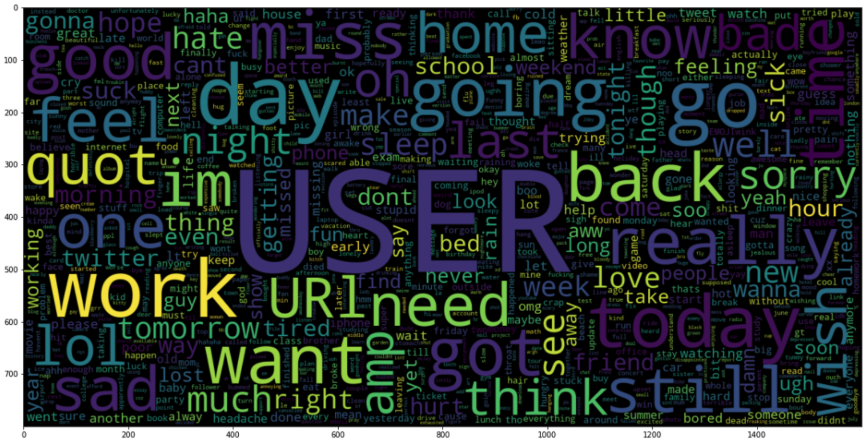


Figure 13: Negative Tweets World Cloud

[7] http://en.wikipedia.org/wiki/List of emoticons
[8] https://www.nltk.org/index.html

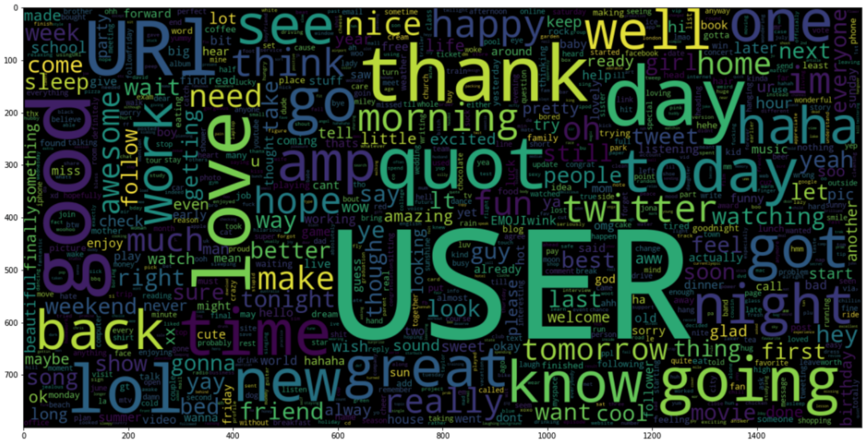
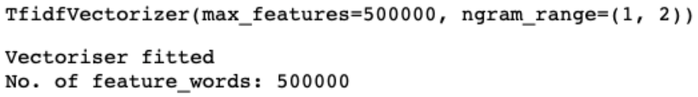


Figure 14: Positive Tweets World Cloud

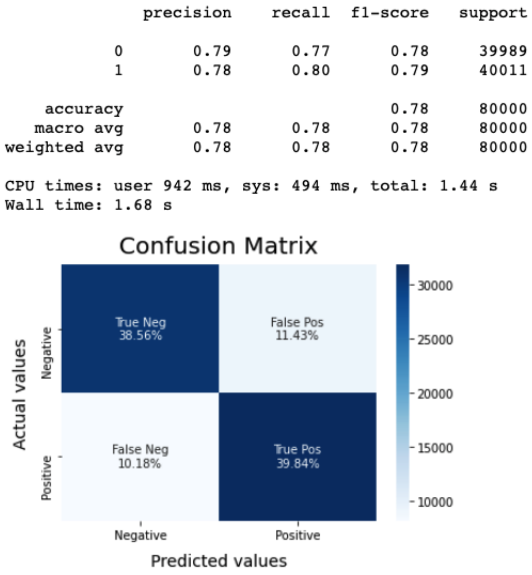
The training and test sets were split into X_train, X_test, y_train, y_test using train_test_split, and then n-gram features were extracted from the tweets and weighted according to TFIDF values. The TFIDF is used to reduce the impact of frequently occurring but less informative tokens, for example, "USER"; the n-gram range is set to (1, 2). The X_train and X_test data sets are then converted to TF-IDF feature matrices using TF-IDF Vectoriser. This data set will be used to train and test the model.



```
TfidfVectorizer(max_features=500000, ngram_range=(1, 2))

Vectoriser fitted
No. of feature_words: 500000
```

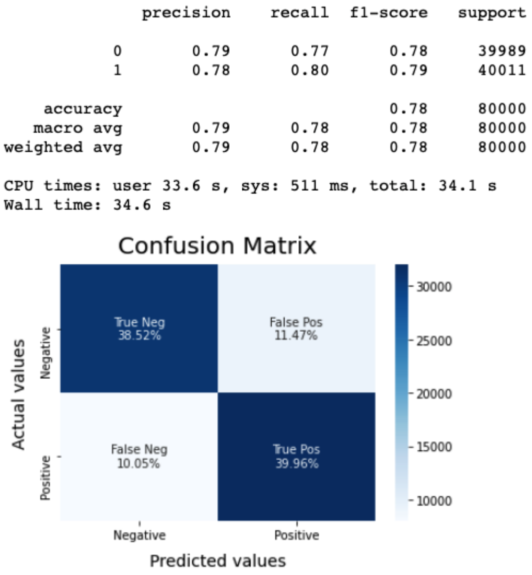
Figure 15: TF-IDF Vectoriser

Three separate models were used to analyse our data set, and a confusion matrix was used to observe the model performance. Accuracy is used as the evaluation metric here.



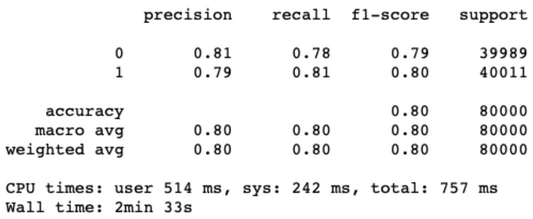
```
              precision    recall  f1-score   support

           0       0.79      0.77      0.78     39989
           1       0.78      0.80      0.79     40011

    accuracy                           0.78     80000
   macro avg       0.78      0.78      0.78     80000
weighted avg       0.78      0.78      0.78     80000

CPU times: user 942 ms, sys: 494 ms, total: 1.44 s
Wall time: 1.68 s
```

Figure 16: Bernoulli Naive Bayes Model



```
              precision    recall  f1-score   support

           0       0.79      0.77      0.78     39989
           1       0.78      0.80      0.79     40011

    accuracy                           0.78     80000
   macro avg       0.79      0.78      0.78     80000
weighted avg       0.79      0.78      0.78     80000

CPU times: user 33.6 s, sys: 511 ms, total: 34.1 s
Wall time: 34.6 s
```

Figure 17: Linear Support Vector Classification Model



```
              precision    recall  f1-score   support

           0       0.81      0.78      0.79     39989
           1       0.79      0.81      0.80     40011

    accuracy                           0.80     80000
   macro avg       0.80      0.80      0.80     80000
weighted avg       0.80      0.80      0.80     80000

CPU times: user 514 ms, sys: 242 ms, total: 757 ms
Wall time: 2min 33s
```

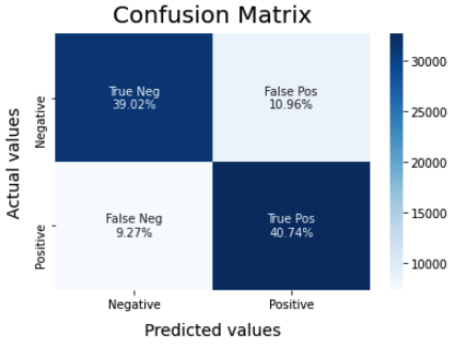
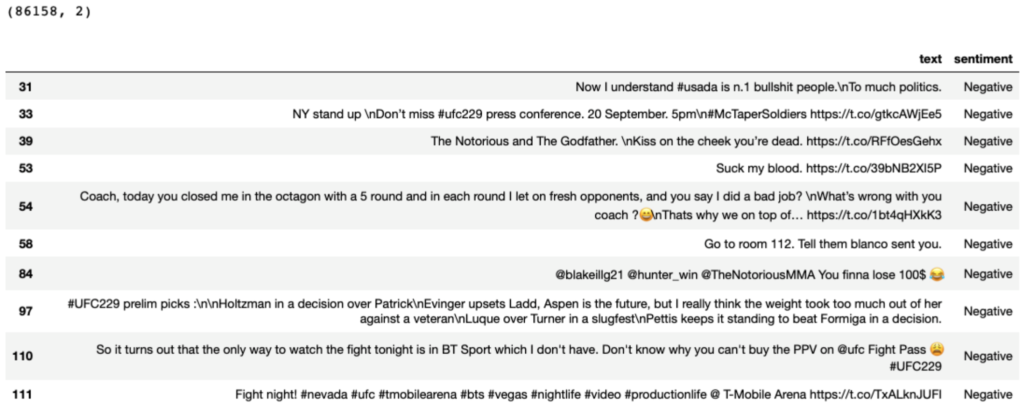


Figure 18: Logistic Regression Model

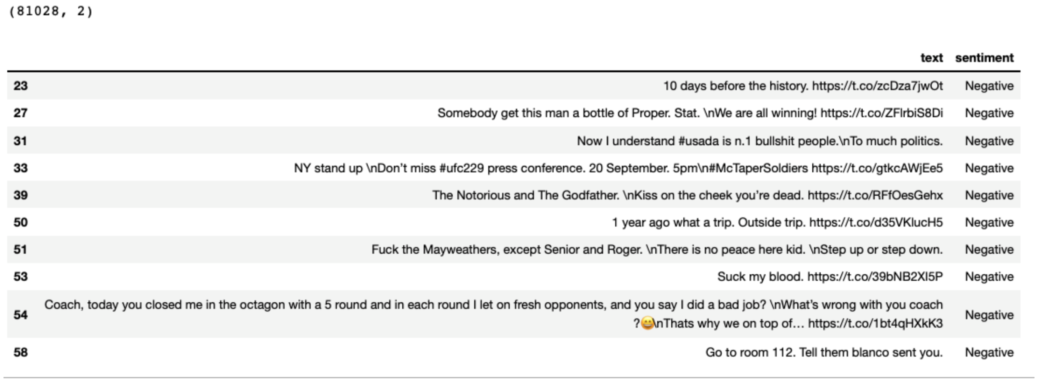
After trying different models, we can observe from the confusion matrix that the Logistic regression model gave the highest accuracy with an f1-score of 80; however, it cannot be ignored that Bernoulli Naive Bayes was the fastest model to train, taking only 1.68 seconds, 90 times faster than the Logistic regression model with an f1- score of 79. Finally, we applied the collected MMA data set to do sentiment classification using Logistic Regression and Bernoulli Naive Bayes models. The result:



```
(86158, 2)
```

	text	sentiment
31	Now I understand #usada is n.1 bullshit people.\nTo much politics.	Negative
33	NY stand up \nDon't miss #ufc229 press conference. 20 September. 5pm\n#McTaperSoldiers https://t.co/gtkcAWjEe5	Negative
39	The Notorious and The Godfather. \nKiss on the cheek you're dead. https://t.co/RFfOesGehx	Negative
53	Suck my blood. https://t.co/39bNB2XI5P	Negative
54	Coach, today you closed me in the octagon with a 5 round and in each round I let on fresh opponents, and you say I did a bad job? \nWhat's wrong with you coach ?😁\nThats why we on top of… https://t.co/1bt4qHXkK3	Negative
58	Go to room 112. Tell them blanco sent you.	Negative
84	@blakeillg21 @hunter_win @TheNotoriousMMA You finna lose 100$ 😂	Negative
97	#UFC229 prelim picks :\n\nHoltzman in a decision over Patrick\nEvinger upsets Ladd, Aspen is the future, but I really think the weight took too much out of her against a veteran\nLuque over Turner in a slugfest\nPettis keeps it standing to beat Formiga in a decision.	Negative
110	So it turns out that the only way to watch the fight tonight is in BT Sport which I don't have. Don't know why you can't buy the PPV on @ufc Fight Pass 😩 #UFC229	Negative
111	Fight night! #nevada #ufc #tmobilearena #bts #vegas #nightlife #video #productionlife @ T-Mobile Arena https://t.co/TxALknJUFI	Negative

Figure 19: MMA Data Set, Negative Words - Logistic Regression Model



```
(81028, 2)
```

	text	sentiment
23	10 days before the history. https://t.co/zcDza7jwOt	Negative
27	Somebody get this man a bottle of Proper. Stat. \nWe are all winning! https://t.co/ZFlrbiS8Di	Negative
31	Now I understand #usada is n.1 bullshit people.\nTo much politics.	Negative
33	NY stand up \nDon't miss #ufc229 press conference. 20 September. 5pm\n#McTaperSoldiers https://t.co/gtkcAWjEe5	Negative
39	The Notorious and The Godfather. \nKiss on the cheek you're dead. https://t.co/RFfOesGehx	Negative
50	1 year ago what a trip. Outside trip. https://t.co/d35VKlucH5	Negative
51	Fuck the Mayweathers, except Senior and Roger. \nThere is no peace here kid. \nStep up or step down.	Negative
53	Suck my blood. https://t.co/39bNB2Xl5P	Negative
54	Coach, today you closed me in the octagon with a 5 round and in each round I let on fresh opponents, and you say I did a bad job? \nWhat's wrong with you coach ?😄\nThats why we on top of… https://t.co/1bt4qHXkK3	Negative
58	Go to room 112. Tell them blanco sent you.	Negative

Figure 20: MMA Data Set, Negative Words - Bernoulli Naive Bayes Model

The results show that 81,028 negative terms were detected using the Bernoulli Naive Bayes model and 86,158 negative terms were detected using the Logistic Regression model.

Since Logistic Regressiony performs relatively well in confusion matrices, this negative data set was sampled and manually labelled. A random sample of 3,000 tweets from 86,158 tweets was selected and manually labelled by three people. Each person obtained the same number and content of tweets, and removed the original sentiment tags to reduce the impact on the test.

5.2 Toxic Language Classification

Due to the excellent training speed and accuracy of Logistic Regression[5], and the fact that this method received the highest score in the Kaggle competition[9], Logistic Regression will continue to be used as a baseline for the classification of toxic languages. The data set for this training is Wikipedia comments data.

As with the previous training, we first observe the data set. Since we are training a model with multiple labels, we use heatmap to observe the correlation between feature and target. There are 6 categories: toxic, severe_toxic, obscene, threat, insult, and identity_hate. The association between insult and obscene is the highest at 68%, while the association between threat and the other categories is very low at less than 0.2.

[9] https://reurl.cc/KQ3lvj

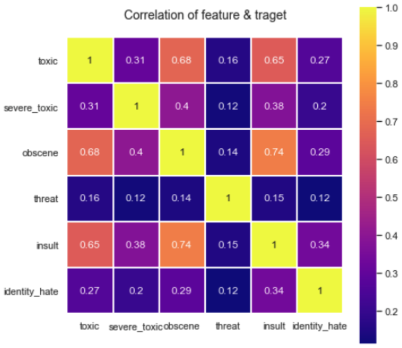
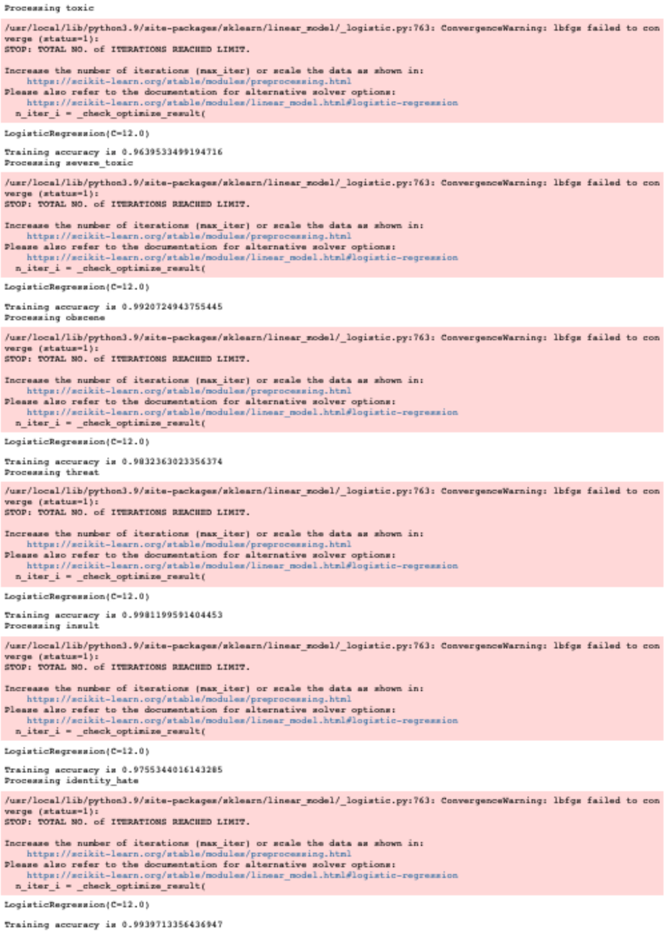


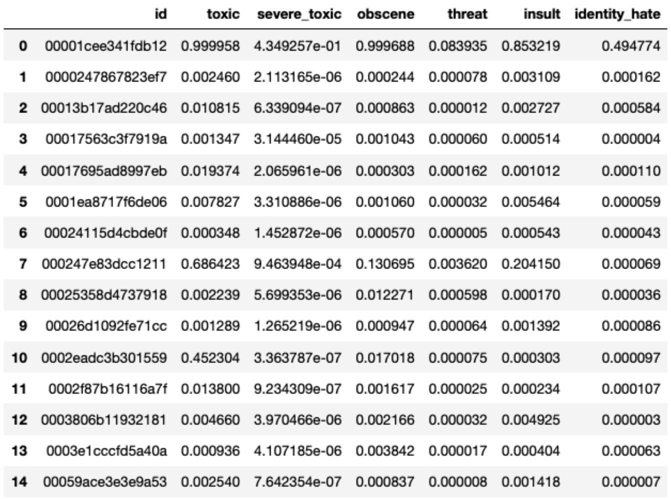
Figure 21: Heatmap for Correlations

As in the previous model, TFIDF was used for weighting, and the weighted data were trained using a logistic regression model. As can be seen from the graphs, the accuracy of the training for each category was over 98%.



```
Processing toxic
/usr/local/lib/python3.9/site-packages/sklearn/linear_model/_logistic.py:763: ConvergenceWarning: lbfgs failed to con
verge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
LogisticRegression(C=12.0)

Training accuracy is 0.9639533499194716
Processing severe_toxic
/usr/local/lib/python3.9/site-packages/sklearn/linear_model/_logistic.py:763: ConvergenceWarning: lbfgs failed to con
verge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
LogisticRegression(C=12.0)

Training accuracy is 0.9920724943755445
Processing obscene
/usr/local/lib/python3.9/site-packages/sklearn/linear_model/_logistic.py:763: ConvergenceWarning: lbfgs failed to con
verge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
LogisticRegression(C=12.0)

Training accuracy is 0.9832363023356374
Processing threat
/usr/local/lib/python3.9/site-packages/sklearn/linear_model/_logistic.py:763: ConvergenceWarning: lbfgs failed to con
verge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
LogisticRegression(C=12.0)

Training accuracy is 0.9981199591404453
Processing insult
/usr/local/lib/python3.9/site-packages/sklearn/linear_model/_logistic.py:763: ConvergenceWarning: lbfgs failed to con
verge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
LogisticRegression(C=12.0)

Training accuracy is 0.9755344016143285
Processing identity_hate
/usr/local/lib/python3.9/site-packages/sklearn/linear_model/_logistic.py:763: ConvergenceWarning: lbfgs failed to con
verge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
LogisticRegression(C=12.0)

Training accuracy is 0.9939713356436947
```

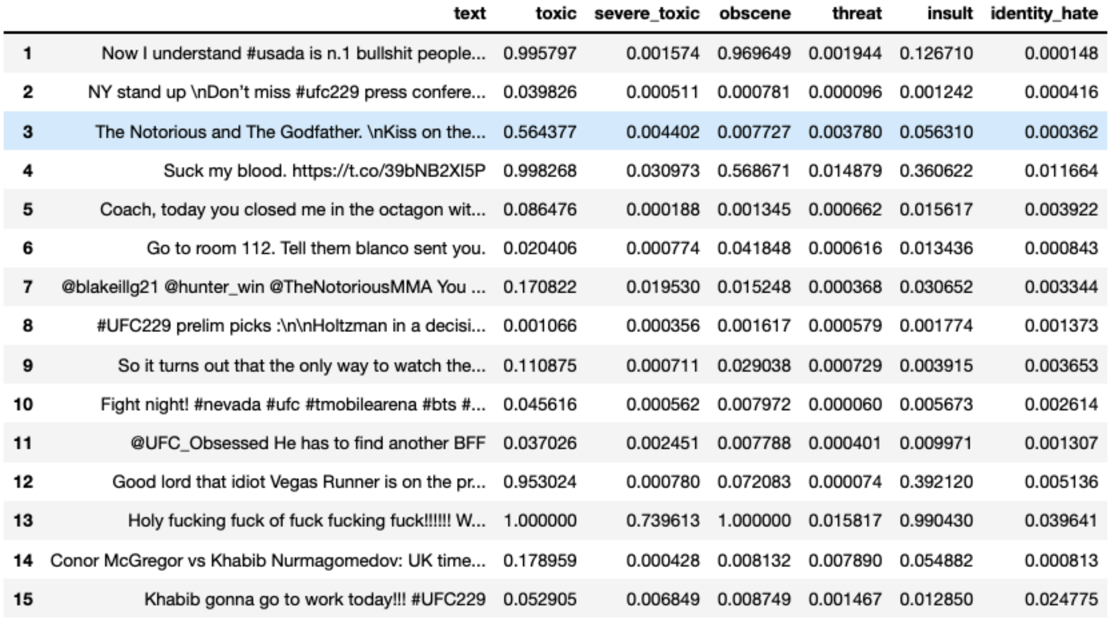
Figure 22: Training Accuracy



	id	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	00001cee341fdb12	0.999958	4.349257e-01	0.999688	0.083935	0.853219	0.494774
1	0000247867823ef7	0.002460	2.113165e-06	0.000244	0.000078	0.003109	0.000162
2	00013b17ad220c46	0.010815	6.339094e-07	0.000863	0.000012	0.002727	0.000584
3	00017563c3f7919a	0.001347	3.144460e-05	0.001043	0.000060	0.000514	0.000004
4	00017695ad8997eb	0.019374	2.065961e-06	0.000303	0.000162	0.001012	0.000110
5	0001ea8717f6de06	0.007827	3.310886e-06	0.001060	0.000032	0.005464	0.000059
6	00024115d4cbde0f	0.000348	1.452872e-06	0.000570	0.000005	0.000543	0.000043
7	000247e83dcc1211	0.686423	9.463948e-04	0.130695	0.003620	0.204150	0.000069
8	00025358d4737918	0.002239	5.699353e-06	0.012271	0.000598	0.000170	0.000036
9	00026d1092fe71cc	0.001289	1.265219e-06	0.000947	0.000064	0.001392	0.000086
10	0002eadc3b301559	0.452304	3.363787e-07	0.017018	0.000075	0.000303	0.000097
11	0002f87b16116a7f	0.013800	9.234309e-07	0.001617	0.000025	0.000234	0.000107
12	0003806b11932181	0.004660	3.970466e-06	0.002166	0.000032	0.004925	0.000003
13	0003e1cccfd5a40a	0.000936	4.107185e-06	0.003842	0.000017	0.000404	0.000063
14	00059ace3e3e9a53	0.002540	7.642354e-07	0.000837	0.000008	0.001418	0.000007

Figure 23: Training Result

Using this model as a baseline, the results were tested using the MMA data set negative vocabulary extracted from the previous model as a test set as shown below.



	text	toxic	severe_toxic	obscene	threat	insult	identity_hate
1	Now I understand #usada is n.1 bullshit people...	0.995797	0.001574	0.969649	0.001944	0.126710	0.000148
2	NY stand up \nDon't miss #ufc229 press confere...	0.039826	0.000511	0.000781	0.000096	0.001242	0.000416
3	The Notorious and The Godfather. \nKiss on the...	0.564377	0.004402	0.007727	0.003780	0.056310	0.000362
4	Suck my blood. https://t.co/39bNB2XI5P	0.998268	0.030973	0.568671	0.014879	0.360622	0.011664
5	Coach, today you closed me in the octagon wit...	0.086476	0.000188	0.001345	0.000662	0.015617	0.003922
6	Go to room 112. Tell them blanco sent you.	0.020406	0.000774	0.041848	0.000616	0.013436	0.000843
7	@blakeillg21 @hunter_win @TheNotoriousMMA You ...	0.170822	0.019530	0.015248	0.000368	0.030652	0.003344
8	#UFC229 prelim picks :\n\nHoltzman in a decisi...	0.001066	0.000356	0.001617	0.000579	0.001774	0.001373
9	So it turns out that the only way to watch the...	0.110875	0.000711	0.029038	0.000729	0.003915	0.003653
10	Fight night! #nevada #ufc #tmobilearena #bts #...	0.045616	0.000562	0.007972	0.000060	0.005673	0.002614
11	@UFC_Obsessed He has to find another BFF	0.037026	0.002451	0.007788	0.000401	0.009971	0.001307
12	Good lord that idiot Vegas Runner is on the pr...	0.953024	0.000780	0.072083	0.000074	0.392120	0.005136
13	Holy fucking fuck of fuck fucking fuck!!!!!! W...	1.000000	0.739613	1.000000	0.015817	0.990430	0.039641
14	Conor McGregor vs Khabib Nurmagomedov: UK time...	0.178959	0.000428	0.008132	0.007890	0.054882	0.000813
15	Khabib gonna go to work today!!! #UFC229	0.052905	0.006849	0.008749	0.001467	0.012850	0.024775

Figure 24: MMA Data Set Result

6 Discussion

The total number of tweets collected in the MMA data set for this project was 338,879, and the total number of negative tweets detected by the model was 86,158, representing only 25% of the entire data set. The results showed a low percentage of toxic comments, with most of the tweets being positive and expressing great anticipation for the race to come. The tweets from both athletes were also not as confrontational and aggressive as they would have been at a public press conference. This fight is not representative of all UFC fights, nor are the two fighters representative of the UFC as a whole, but this fight and the two fighters are among the best known fighters in the UFC.

7 Conclusion

Based on analysis and categorisation, mixed martial arts fighters are far less toxic in their tweets than in their press conferences[10]. However, the results of the analysis carried out for this project do not mean that the same can be said for the statements and actions of the fighters on other platforms. Conor clashed with Kabib's team before the fight was confirmed, and was even sued for attacking a bus[11].

The end result of the project showed that the toxic or hateful comments used by fighters and fans on Twitter were not as bad as initially expected.

8 Future Work

Due to time and environmental constraints, it was not possible to collect more tweets and use a better but longer model (LSTM with word embedding) for the analysis. In the future, it might be more representative and accurate if we could collect tweets from the players' daily posts or tweets about potential opponents on Twitter before the official signing of the tournament.

References

[1] Roland Barthes. *Mythologies*. Paris: Éditions du Seuil, 1957.

[2] Nhan Cach Dang, Marıa N Moreno-Garcıa, and Fernando De la Prieta. "Sentiment analysis based on deep learning: A comparative study". In: *Electronics* 9.3 (2020), p. 483.

[3] Kerstin Denecke. "Using sentiwordnet for multilingual sentiment analysis". In: *2008 IEEE 24th international conference on data engineering workshop*. IEEE. 2008, pp. 507–512.

[4] Björn Gambäck and Utpal Kumar Sikdar. "Using convolutional neural networks to classify hate-speech". In: *Proceedings of the first workshop on abusive language online*. 2017, pp. 85–90.

[5] Aditya Gaydhani et al. "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach". In: *arXiv preprint arXiv:1809.08651* (2018).

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

[10] https://www.youtube.com/watch?v=agbsUzk3gwct=1450sab$_channel = BTSport$

[11] https://www.youtube.com/watch?v=Xc56YrGbwu0ab$_channel = InsightUp$

[7] Alex Graves and Jürgen Schmidhuber. “Framewise phoneme classification with bidirectional LSTM and other neural network architectures”. In: *Neural networks* 18.5-6 (2005), pp. 602–610.

[8] Ramish Jamil et al. “Detecting sarcasm in multi-domain datasets using convolutional neural networks and long short term memory network model”. In: *PeerJ Computer Science* 7 (2021), e645.

[9] Subbu Kannan et al. “Preprocessing techniques for text mining”. In: *International Journal of Computer Science & Communication Networks* 5.1 (2014), pp. 7–16.

[10] David G Kleinbaum et al. *Logistic regression.* Springer, 2002.

[11] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. “Twitter sentiment analysis: The good the bad and the omg!” In: *Proceedings of the international AAAI conference on web and social media.* Vol. 5. 1. 2011, pp. 538–541.

[12] Matt Kusner et al. “From word embeddings to document distances”. In: *International conference on machine learning.* PMLR. 2015, pp. 957–966.

[13] Scott Menard. *Applied logistic regression analysis.* 106. Sage, 2002.

[14] Tomas Mikolov et al. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems* 26 (2013).

[15] Zhu Nanli et al. “Sentiment analysis: A literature review”. In: *2012 International Symposium on Management of Technology (ISMOT).* IEEE. 2012, pp. 572–576.

[16] Juan Ramos et al. “Using tf-idf to determine word relevance in document queries”. In: *Proceedings of the first instructional conference on machine learning.* Vol. 242. 1. Citeseer. 2003, pp. 29–48.

[17] Yafeng Ren et al. “Context-sensitive twitter sentiment classification using neural network”. In: *Thirtieth AAAI conference on artificial intelligence.* 2016.

[18] Jenq-Haur Wang et al. “An LSTM approach to short text sentiment classification with word embeddings”. In: *Proceedings of the 30th conference on computational linguistics and speech processing (ROCLING 2018).* 2018, pp. 214–223.

[19] Sida I Wang and Christopher D Manning. “Baselines and bigrams: Simple, good sentiment and topic classification”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* 2012, pp. 90–94.

[20] Lei Zhang, Shuai Wang, and Bing Liu. “Deep learning for sentiment analysis: A survey”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4 (2018), e1253.