Sentiment Analysis in Twitter Using Machine Learning

Bahi Ibrahim Elsayed Mohamed
Obour Institute Computer Science
Bahi.Ibrahim_617032@oi.edu.eg

Mohammed Ahmed El-Sayed
Obour Institute Computer Science
Mohamed_617014@oi.edu.eg

Rasha elStohy
Computer Science Lecturer
rashastohy@oi.edu.eg

Abstract: **Social networks, such as Twitter, are the most popular applications for exchanging ideas on various topics or events. As a result, these types of applications provide a rich supply of data for Natural Language Processing (NLP) researchers to develop and improve techniques for tracking users' sentiments about a specific event, issue, or even another person. These users' attitudes are important for decision makers since they can take appropriate action in response to users' negative or positive emotions. This research focuses on detecting users' attitudes using a new feature set and applying several machine learning models to monitor and improve users' attitude detection systems. In order to train, create, and test classification models, the dataset of annotated emotion tweets and the word emotion lexicon are used.**

Keyword: NLP,Twistter,Social networks,Machine Learning

I. INTRODUCTION

Sentiment analysis offers a wide range of applications, including obtaining and analysing individual opinions on a variety of items, issues, social, and political events. Understanding public opinion can help you make better decisions [1]. Opinion mining is a method of gathering data from search engines, blogs, microblogs, and social media sites. Individual ideas differ from person to person, and Twitter tweets are an excellent source of this type of information. However, assessing text/opinion data efficiently is difficult because to the large volume and unstructured nature of the data. . Sentiment analysis is a Natural Language Processing (NLP) activity that involves detecting and categorising sentiments in texts. The “positive,” “negative,” and “neutral” classifications are typically considered [3].

As a result, skilled algorithms/computational methodologies are necessary for data mining and condensing, as well as discovering sentiment or even finer-grained emotions in words. Different machine learning techniques are used to determine the efficiency of various feature combinations as well as to identify the sentiment of tweets. [4]. The supervised strategy was utilised in this study to incorporate unigram, bigram, and Part-of-Speech as features, which is effective in identifying emotion and sentiment in unstructured data.application standpoint in computational linguistics. The distinction between opinion, sentiment, and emotion is a little ambiguous. Opinion is a transitory concept that expresses a person's opinion regarding something. Emotion represents attitude, whereas sentiment reflects feeling [5].

Many Scientists disagree about the existence of more basic emotions than others, as well as about what and how many basic emotions are, but the most common emotions in emotion detection methods so far are listed in the table [6] below.At this paper we tried to use machine leraning techniques to extract emotion for teenths and youth specially in lockdown duration to rech if it affects their tweets or not.

II. RELATED WORKS

Instead of using the traditional methods of sentiment analysis of positive and negative analysis, it can be more precise and specific such as introducing multiple emotions to improve and

strengthen the classification.This can be achieved through the use of multiple cognitive and lexical bases and an appropriate approach to dealing with this type of classification problem (text classification) and it also has the ability to deal with natural

Table 1: The most used emotions in emotion detection.

Lists of Basic Emotions	
Ekman	anger, disgust, fear, joy, sadness, and surprise
Pultchick	anger, anticipation, disgust, fear, joy, sadness, surprise, and trust
Izard	anger, Contempt, disgust, distress, fear, guilt, interest, joy, shame, and surprise

the machine learning approaches used to classify tweets.[10]

At our research, we utilized Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models.

Multilayer Propon (MLP), Nave Bayes, Fuzzy Classification, Decision Tree, and Support Vector Machines (SVM) are among

III. GENERAL FRAMEWORK ARCHITECTURE

Twitter is a social networking service that allows broadcasting of short messages called tweets. These tweets from millions of active users all over the world are considered as an information treasure, that not only attracts academics attention to know what users' interests are but also organizations as well language preprocessing as well as machine learning [7-9].

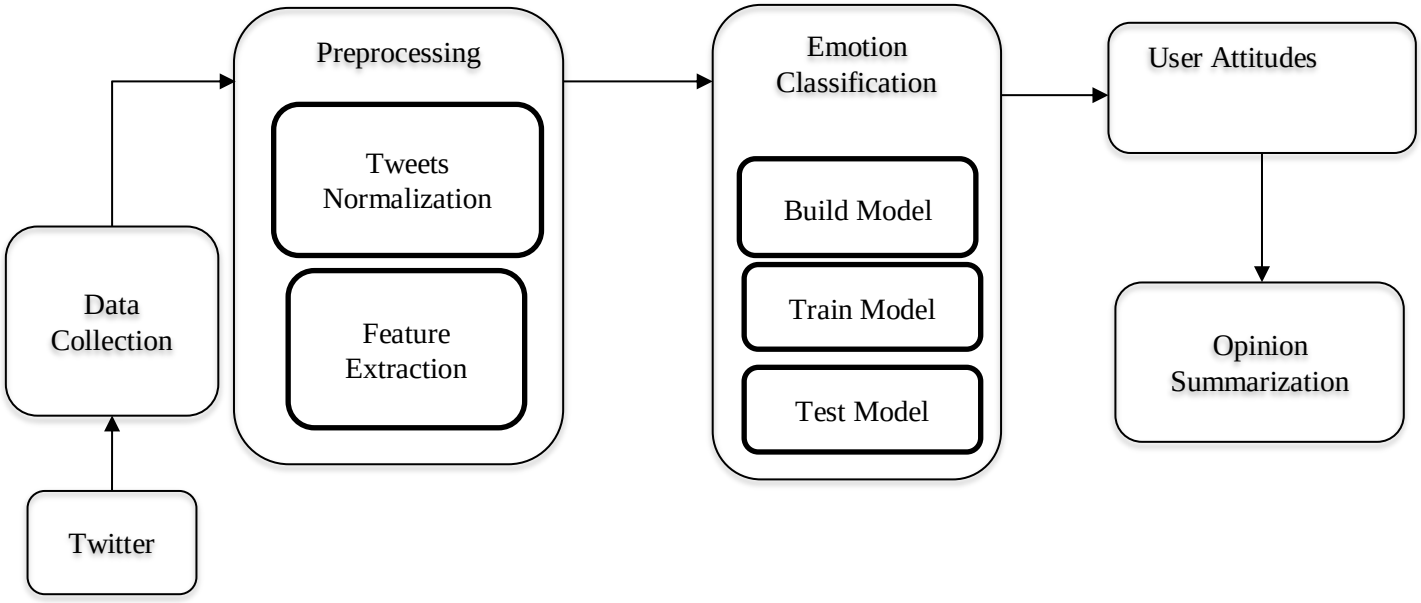

Figure 1: Emotion Detection Workflow

Fig. 1 shows the general users' attitude detection workflow. It shows sequence of processes in supervised machine learning algorithms in sentiment analysis and emotion detection. It begins with data collection which concerns to collect relative tweets from Twitter.

A)Data Preprocessing

In your workflow, data preparation is quite crucial. You must change the data in such a way that it can be processed by a machine. A database is nothing more than a collection of data elements. Data samples, events, observations, and records are all terms that can be used to describe them. However, distinct traits are used to describe each of them.

They're referred to as attributes or features in data science jargon.Data preprocessing is a necessary step before building a model with these features.

preprocessing and feature extraction phase. It operates on filtering significant tokens, and encodes them to a proper feature vector. Then, emotion classification phase concerns about model building, training, and testing. Finally, opinion summarization phase collects and threshold output to generate final decision to the sample input.

Machine learning process steps in Data Preprocessing

Step 1: Import the libraries

Step 2: Import the data-set

Step 3: Splitting the data-set into Training and Test Set

Step 4: translate emojis

Step 5: remove urls

Step 6: remove stopwords & punctuations & unneeded spaces

	anger	anticipation	disgust	fear	joy	love	optimism	pessimism	sadness	surprise	trust
0	2544	978	2602	1242	2477	700	1984	795	2008	361	357

Fig 2: Counting each emotion

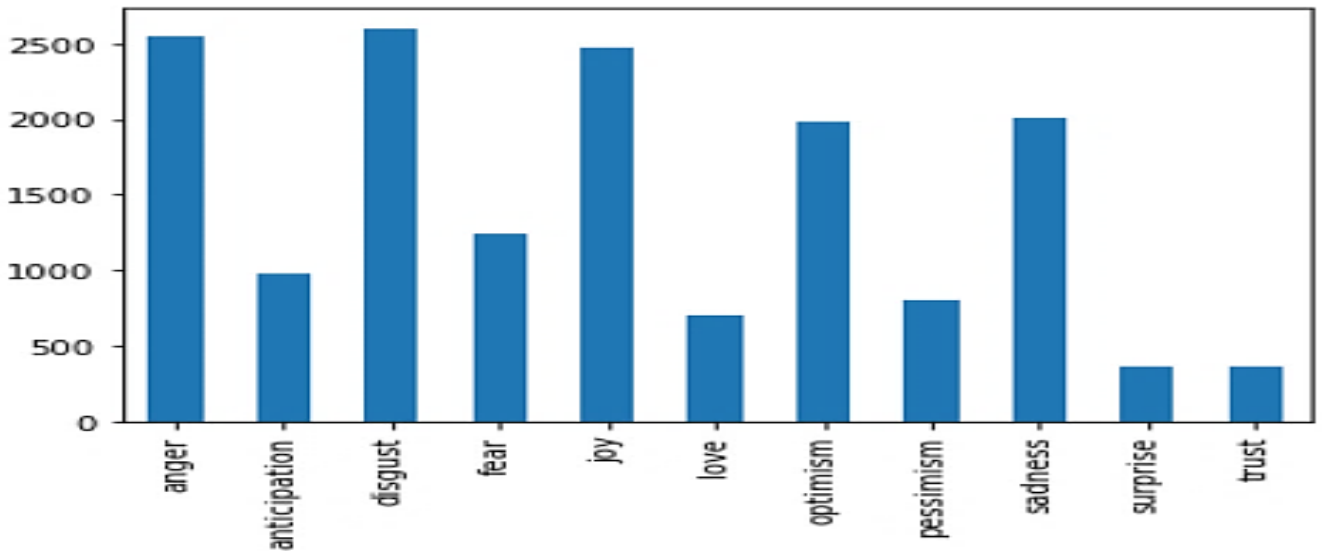

Fig 3: Counting each emotion

B.Feature Extraction

Feature extraction is a dimensionality reduction procedure that reduces a large set of raw data into smaller groupings for processing. The enormous number of variables in these large data sets necessitates a lot of computational resources to process. Feature extraction refers to strategies for selecting and/or combining variables into features in order to reduce the amount of data that needs to be processed while still accurately and thoroughly characterising the original dataset[12].

C.Language modeling

The specialty of language modelling is determining the likelihood of a sequence of words. These are useful in a variety of NLP applications, including machine translation, speech recognition, optical character recognition, and many others. In recent years, language models have relied on neural networks to predict a word in a phrase based on surrounding terms. [13].

In this work will discuss the most classic of language models: the **n-gram** models. In natural language processing, an n-gram is an arrangement of n words

What are N-grams (unigram, bigram)?

An N-gram is a sequence of N tokens (or words).

Let's understand N-gram with an example. Consider the following sentence:

"I like reading books about data science while drinking tea."

A unigram is a one-word sequence. For the above sentence, the unigrams would simply be: "I", "like", "reading", "books", "about", "data","science", "while", "drinking", "tea".

A bigram is a two-word sequence of words, like "I like", "like reading", or "drinking tea".

A frequency distribution is a table that shows how many times each word appears in a piece of text. Frequency distributions are a special object type in NLTK, and they are implemented as a separate class named FreqDist. This class contains procedures that are useful for word frequency analysis, sample of Unigram frequency distribution and Explanatory Analysis ass in fig4, fig5 demonstrate Bigram frequency distribution and Explanatory Analysis

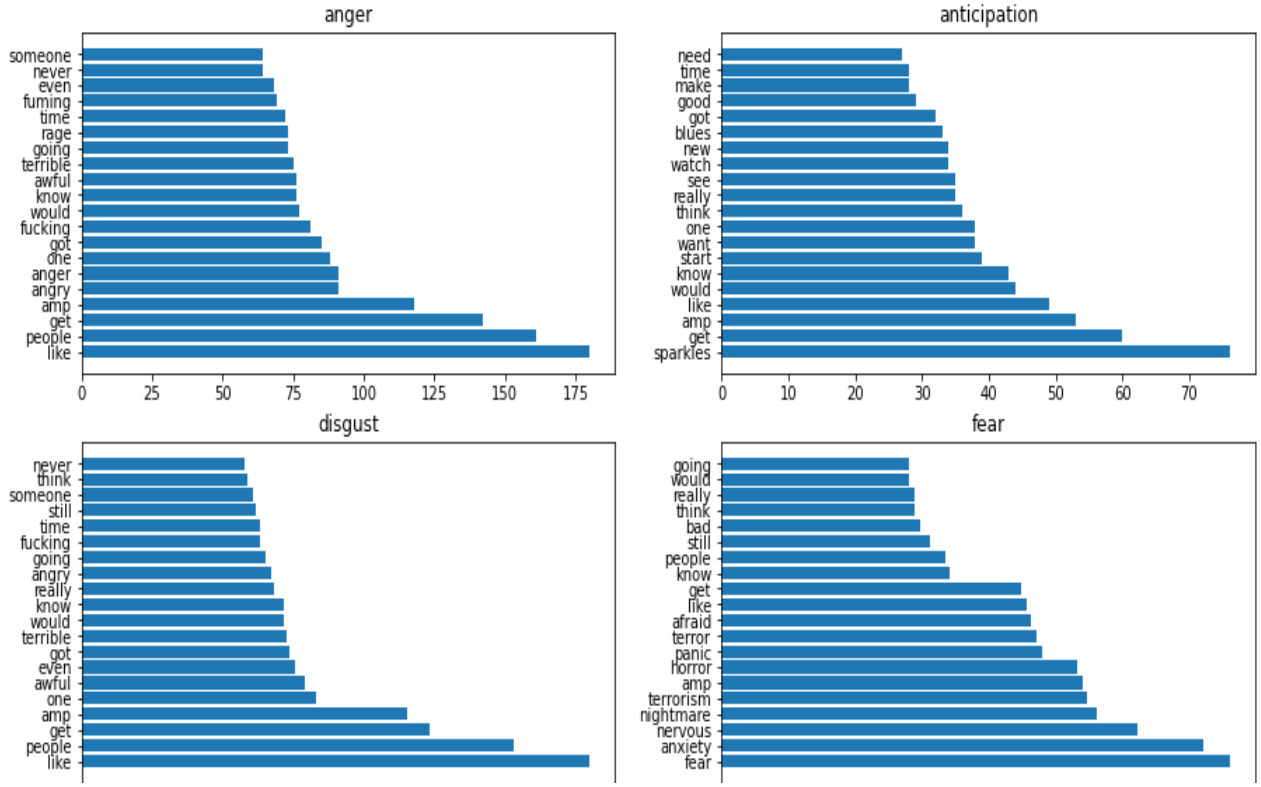

Fig4. sample of Unigram frequency distribution and Explanatory Analysis

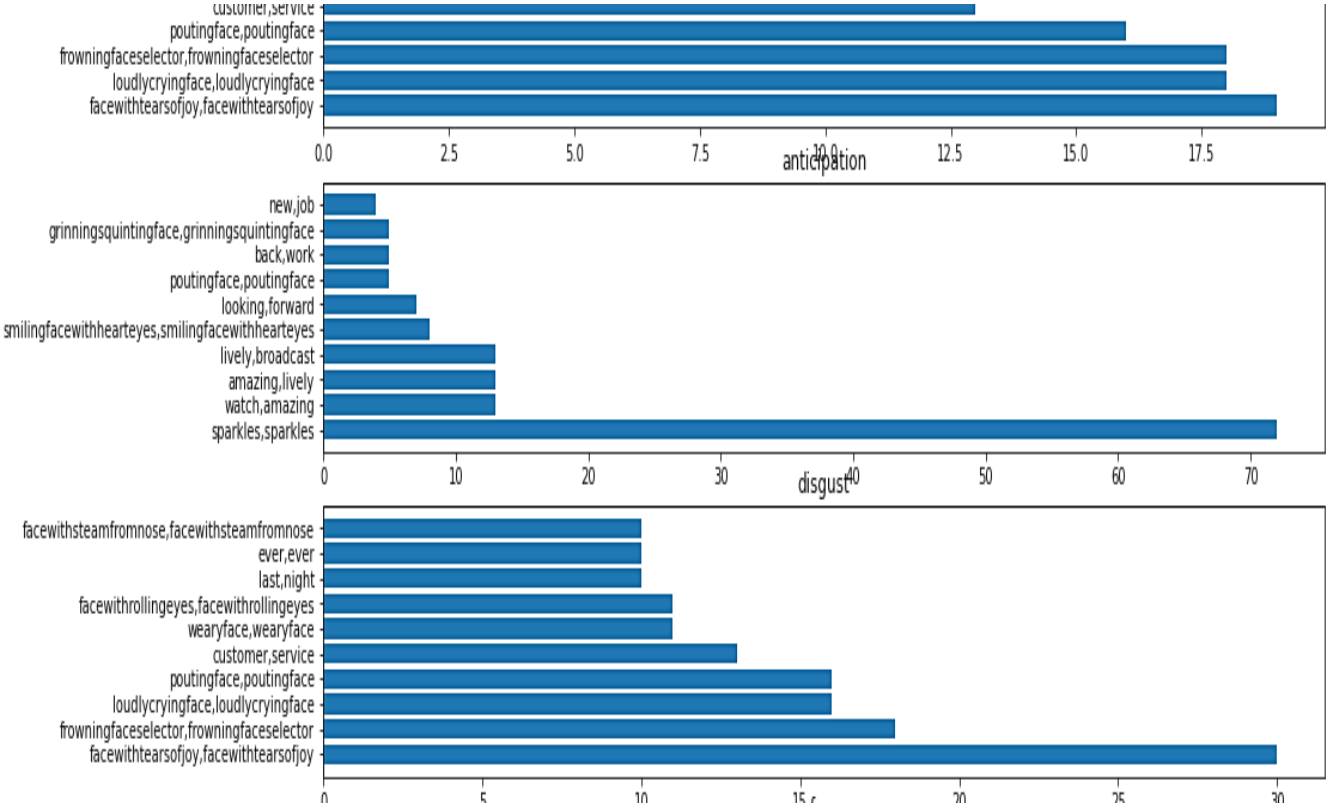

Fig5. Bigram frequency distribution and Explanatory Analysis

IV.Lexicons

A.Word lexicon

A lexicon is a collection, definition, and explanation tool for a language's vocabulary [14].

This research focuses on the problem of disambiguating polarity-ambiguous words, and it reduces the task to sentiment classification of aspects, which it refers to as sentiment expectation rather than semantic orientation, which has been frequently employed in prior studies. Words like "huge, little, high, low" are examples of polarity-ambiguous words, which make sentiment analysis difficult. This work uses a mutual bootstrapping technique to generate the aspect and polarity-ambiguous lexicon in order to disambiguate polarity-ambiguous words. As a result, the sentiment of polarity-ambiguous words in context can be determined jointly based on the sentiment expectations of the aspects and the antecedent polarity of the polarity-ambiguous words. Experiments reveal that its method is effective in sentiment analysis at the sentence level[15].

NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions and two sentiments. The annotations were manually done by crowdsourcing [16].

Sentiment analysis can also be expanded by categorizing the evaluations based on the predominant emotion. These more recent advances are based on the N.R.C. lexicon, which contains 14,000 words and defines the emotional sentiment in words in great detail. National as Research Council (NRC) Emotion Lexicon can Comparing preprocessed words with emotion word lexicon, to detect or convey sentiments can be used as key features. The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive) [17 - 20]; as well as the recognition of the most common terms associated with each emotion.

The NRC Emotion Lexicon contains affect annotations on English words. Despite certain geographical differences, it has been discovered that the diversity of affective norms is consistent across languages. The dictionary was made available in over a hundred languages by utilising Google Translate to translate the English words .

Felipe Bravo-Marquez developed Affective Tweets, a collection of filters for extracting features from tweets for emotion classification/regression and other activities, for the Weka machine learning workbench. The package comes in handy when you need to build function vectors from a wide number of effect lexicons. To increase performance, the vector may be concatenated with other features vectors (for example, dense-distributed representations of the text). (The function vector can be used for any classifier, not just those that support Weka).

B.Hash Lexicon

This method necessitates the creation of a dictionary of positive and negative words, each having a positive or negative emotion value ascribed to it. A chunk of text communication is typically represented as a bag of words in lexicon-based techniques. All positive and negative words or phrases inside the message are

One of the two primary techniques to sentiment anal

given emotion values from the lexicon. To make the final prediction about the overall sentiment for the message, a combining function, such as sum or average, is used. Aspects of a word's local context, such as negation or intensification, are frequently taken into account in addition to its sentiment value.

We chose to use a lexicon-based strategy in our research to avoid having to create a tagged training set. Machine learning algorithms have a major drawback in that they rely on tagged data. We use computers and algorithms to help us sift through massive volumes of data and interpret what we're seeing.

The NRC has developed Sentiment and Emotion Lexicons, which enable the design of software for automatic sentiment analysis, enabling a deeper understanding of the underlying feelings and emotions included within information, based on advanced experience in Text Analytics.

The lexicons have many uses, including:

• Improving customer relation models

• Identifying what evokes strong emotions in people

• Tracking sentiment towards politicians, movies, products

• Detecting happiness and well-being

• Improving automatic dialogue and tutoring systems

• Detect how emotional words and metaphors are used to persuade and constrain

• Developing affect-sensitive characters in computer games.

We utilized Word-Emotion Association (a.k.a. NRC Emotion Lexicon which Lists associations of words with eight emotions

Table 2: Summary Details of the NRC Emotion Lexicon [21]

Association Lexicon	*Version*	*# of Terms*	*Categories*	*Association Scores*	*Method of Creation*	*Papers*
NRC Word-Emotion Association Lexicon (also called EmoLex)	*0.92* *(2010)*	*14,182 unigrams (words)* *~25,000 senses**	*sentiments: negative, positive emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust*	*0 (not associated) or 1 (associated)* *not associated, weakly, moderately, or strongly associated*	*Manual:* *By crowdsourcing on Mechanical Turk.* *Domain:* *General*	*Crowdsourcing a Word-Emotion Association Lexicon, Saif Mohammad and Peter Turney, Computational Intelligence, 29 (3), 436-465, 2013.* *Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon, Saif Mohammad and Peter Turney, In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, June 2010, LA, California.*

The lexicons have many uses, including:

Improving customer relation models,Identifying what evokes strong emotions in people,Tracking sentiment towards politicians, movies, products,Detecting happiness and well-being,Improving automatic dialogue and tutoring systems and Detect how emotional words and metaphors are used to persuade and constrain Developing affect-sensitive characters in

Hashtag Sentiment: Lists associations of words with positive (negative) sentiment. Generated automatically from tweets with sentiment-bearing hash tagged words such as #amazing and #terrible.

Hashtag Affirmative Context Sentiment and Hashtag Negated Context Sentiment Lists associations of words with positive (negative) sentiment in games affirmative or negated contexts.

C.The NRC Sentiment and Emotion Lexicons included in this distribution:

Word-Emotion Association (a.k.a. NRC Emotion Lexicon): Lists associations of words with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). Created by manual annotation on a crowdsourcing platform

D.VAD Lexicon

Words play a central role in language and thought. Several effective factor analysis studies have shown that the primary dimensions of word meaning are valence, arousal, and dominance (VAD) [22].

Valence is the positive--negative or pleasure--displeasure dimension;Arousal is the excited--calm or active--passive dimension; Dominance is the powerful--weak or 'have full control'--'have no control' dimension.

The NRC Emotion Lexicon is a list of English words and their associations with eleven basic emotions (anger, anticipation, disgust, fear, joy, love),NRC Valence, Arousal, and Dominance (VAD) optimism, pessimism, sadness, surprise, trust) and three sentiments (negative, positive, and neutral). The Lexicon includes a list of more than 20,000 dominance scores. For a given word and a dimension (V/A/D), the scores range from 0 (lowest V/A/D) to 1 (highest V/A/D). The lexicon with its Very accurate real-valued scores was created by manual annotation using Best--Worst Scaling. The lexicon is markedly larger than any of the existing VAD lexicons. We also show that the ratings obtained are substantially more reliable than those in existing lexicons [23].

Table 3: Entries with Highest and Lowest Scores in the VAD

		Highest			Lowest	
Dimension		word	score		word	score
valence		*love*	1.000		*toxic*	0.008
		happy	1.000		*nightmare*	0.005
arousal		*abduction*	0.990		*mellow*	0.069
		exorcism	0.980		*siesta*	0.046
dominance		*powerful*	0.991		*frail*	0.069
		leadership	0.981		*weak*	0.045

Table 4: Average valence, arousal, and dominance scores

Emotion	Avg. Valence	Avg. Arousal	Avg. Dominance
anger	0.26	0.66	0.46
fear	0.29	0.66	0.48
joy	0.77	0.52	0.60
sadness	0.24	0.58	0.38

Average valence, arousal, and dominance scores for each basic emotion, the cells are in shades of green with the darkness proportional to the score: lighter shades indicate low scores and darker shades indicate high scores [24].

IV. RESULTS AND DISCUSSION

A) At our experiment we utilized The NRC VAD Lexicon that's it has a broad range of applications in Computational Lingustics, Psychology, Digital Humanities, Computational Social Sciences, and beyond. Notably it can be used to:

1. Study how people use words to convey emotions.

2. Study how different genders and personality traits impact how we view the world around us.

3. Study how emotions are conveyed through literature, stories, and characters.

4. Obtain features for machine learning systems in sentiment, emotion, and other affect-related tasks and to create emotion-aware word Embeddings and emotion-aware sentence representations.

5. Evaluate automatic methods of determining V, A, and D.

6. Study the interplay between the basic emotion model and the

VAD model of emotions [25].

7. Study the role of high VAD words in high emotion intensity sentences, tweets, snippets from literature.

B)Machine Learning Algorithm

At our experiment, we utilized Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions .It is a probabilistic classifier, which of used algorithm illustrated at Fig6. means it predicts spam filtration, Sentimental analysis, and classifying articles on the basis of the probability of an object.

Table 5: An Interactive Visualization of the NRC VAD

term	valence	arousal	dominance
aaaaaaah	0.479	0.606	0.291
aaaah	0.520	0.636	0.282
aardvark	0.427	0.490	0.437
aback	0.385	0.407	0.288
abacus	0.510	0.276	0.485
abalone	0.500	0.480	0.412
abandon	0.052	0.519	0.245
abandoned	0.046	0.481	0.130
abandonment	0.128	0.430	0.202
abashed	0.177	0.644	0.307
abate	0.255	0.696	0.604
abatement	0.388	0.338	0.336
abba	0.562	0.500	0.480
abbey	0.580	0.367	0.444
abbot	0.427	0.321	0.483
abbreviate	0.531	0.375	0.330
abbreviation	0.469	0.306	0.345
abdomen	0.469	0.462	0.471

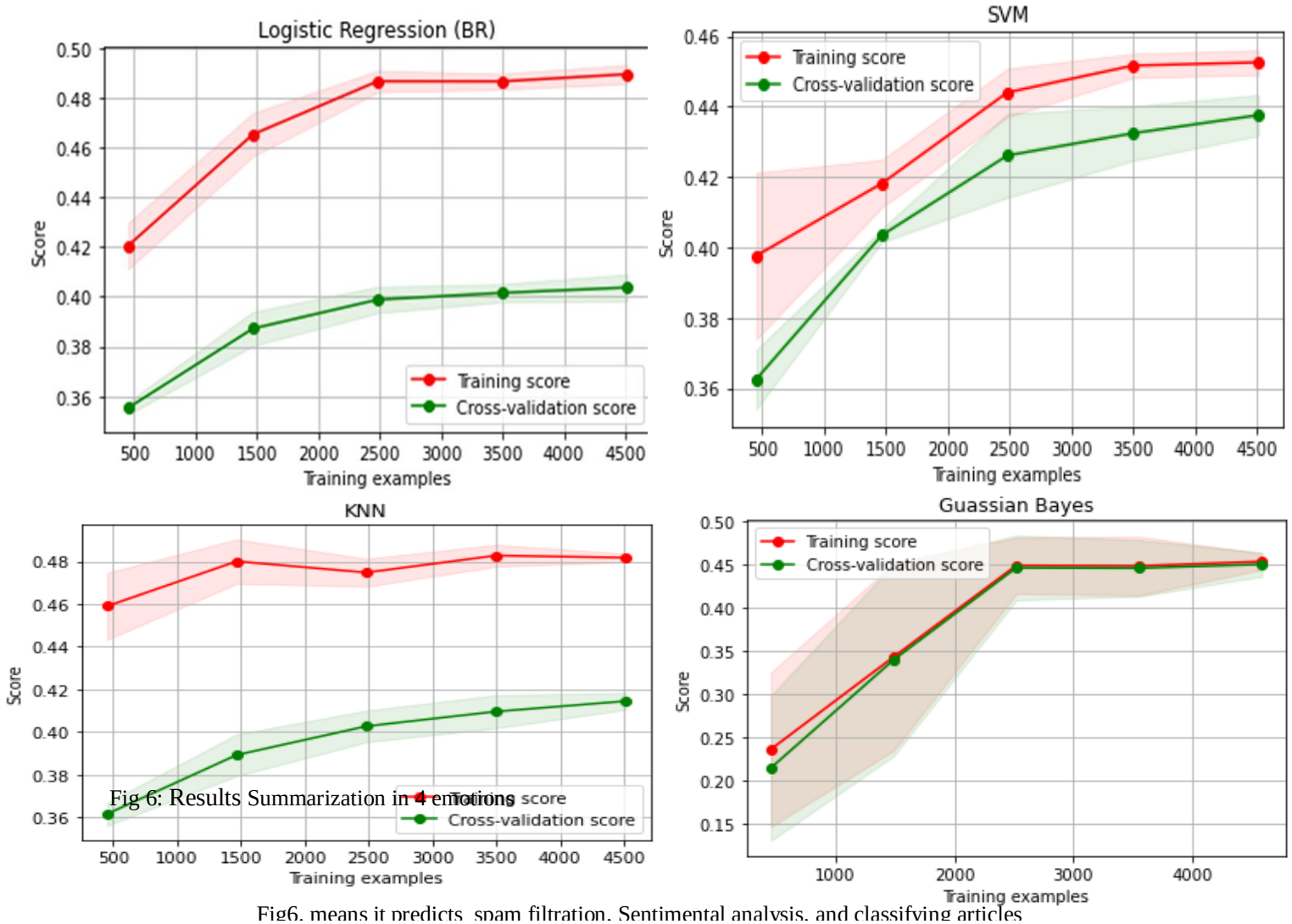

Fig 6: Results Summarization in 4 emotions

Fig6. means it predicts, spam filtration, Sentimental analysis, and classifying articles

V. Conclusion

Based on artificial intelligence and machine learning, this work deals in new approach to extract opinion and analysis sentiment in multi label classification based on real life tweets database classified it into various emotional categories.

Feature vectors were used to represent the tweets, overcoming obstacles such as unstructured slang and bipolarity language. Different classifiers architectures were built and trained them over real-world data as it produced different results (because of feature vector, and the target emotion range) Depending on the algorithm which were used in this Work on this project showed that the experiment was carried out on a scale of four, eight and eleven (our main work) emotions. The results showed that four emotion range was better than eight emotion range.as that eight emotion classifier still need more tweets to train on.

Turning to the classifiers side, in the four-emotion range, a single convolutional neural network and binary relevance performed better in multi emotion detection. However, if the architecture of the convolutional neural network is improved, it can improve its efficiency even further. Both feature vectors generated satisfactory results in the four emotion range detection and were favored in the feature vectors perspective.

REFERENCES

1. Bikel, D. M., & Sorensen, J. (2007). If we want your opinion. In: International conference on semantic computing (ICSC 2007) pp. 493–500.
2. Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2),pp.15–21.
3. Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach toopinion mining. In: Proceedings of the 2008 international conference on web search and data mining, pp. 231– 240.
4. Chen, R., & Xu, W. (2016). The determinants of online customer ratings: A combined domain ontology and topic text analytics approach. Electronic Commerce Research.
5. Adwan omar& Marwan el-tawil & Ammar M huneiti & Rawan shahin(2020),."Twitter Sentiment Analysis Approaches:A Survey". International Journal of Emerging Technologies in Learning (iJET)pp.1-8.
6. Samal, B., Behera, A. K., & Panda, M. (2017). Performance analysis of supervised machine learning techniquesfor sentiment analysis, Third International Conference on Sensing, Signal Processing and Security (ICSSS),pp1-8.
7. Hassan, A. U., Hussain, J., Hussain, M., Sadiq, M., & Lee, S. (2017), Sentiment Analysis of Social Networking Sites (SNS) Data using Machine Learning Approach for the Measurement of Depression. International Conference on Information and Communication Technology Convergence (ICTC)
8. Schukla, A., "Sentiment analysis of document based on annotation",(2018) CORR Journal, Vol. abs/1111.1648.
9. . Mikalai Tsytsarau, Themis Palpanas," Survey on mining subjective data on the web Data Min Knowledge Discovery(2012), vol (24) pp. 478-514
10. Walaa Medhat, Ahmed Hassan, Hoda Korashy, " Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal(2014), Vol 5, Issue 4, pp. 1093-1113.
11. Mohamed Haggag, Samar Fathy, Nahla Elhaggar, "Ontology-Based Textual Emotion Detection" (2015), (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 9, pp. 239- 246.
12. Crowdsourcing a word-emotion association lexicon. Computational Intelligence(2013), 29(3):436– 465.
13. Bryan Orme. (2009), Maxdiff analysis: Simple counting, individual-level logit, and HB. Sawtooth Software, Inc.
14. Mohammad, (2018) LREC paper.
15. Saif M. Mohammad. (2020), Sentiment analysis: Detecting valence, emotions, and other affectual states from text.
16. Lambov Dinko, Pais Sebastião, Dias Gãel(2011). Merged agreement algorithms for domain independent sentiment analysis. In: Presented at the Pacific Association for, Computational Linguistics (PACLING'11).
17. Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly availablelexical resource for opinion mining In: Proceedings of 5th language resources and evaluation Vol. 6, pp. 417–422.
18. Van de Kauter, M., Breesch, D., & Hoste, V. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. ExpertSystems with Applications, 42(11), 4999–5010.
19. Alvaro Ortigosa, Jose M. Martin, Rosa M. Carro(2014).
20. Yeow, R. Mahmud and R. G. Raj, "An application of case-based reasoning with machine learning for forensic autopsy"(2014), Expert Syst. Appl.,vol. 41, no. 7, pp. 3497-3505,.
21. Optimism of early AI(1993),Herbert Simon quote: Simon 1965, p. 96
22. Kaiquan Xu, Stephen Shaoyi Liao, Jiexun Li, Yuxia Song

23.zhou L, Li B, Gao W, Wei Z, Wong K.(2011) Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In: Presented at the 2001 conference on Empirical Methods in Natural Language Processing (EMNLP'11).

24. Heerschop B, Goossen F, Hogenboom A, Frasincar F, Kaymak U, de Jong F.(2011), Polarity Analysis of Texts using Discourse Structure. In: Presented at the 20th ACM Conference on Information and Knowledge Management (CIKM'11).
25. Jonathan Ortigosa-Hernández, Juan Diego Rodríguez, Leandro Alzate, Manuel Lucania, Iñaki Inza, Jose A. Lozano