Sarcasm Detection using Emoji and Text analysis

Fall 2018 CSE 472 Social Media Mining Project II (Group 10)

Jayashree Subramanian (1214347796) jsubram5@asu.edu
Varun Sridharan (1215146660) vsridh19@asu.edu

I. INTRODUCTION

The growth of social media has been exponential and large amount of data is being used in the social media. This huge publicly available data can be used for research and a variety of applications. Recently, the usage of emojis in social media has become more popular. This raises the question of whether these emojis will come to replace earlier methods of paralinguistic communication. Identification of sarcastic comments can reduce misleading data mining activities and reduces the result obtained by wrong classification.

However, it is also used to determine how emoji can be influential in social media and if we can detect sarcasm from emojis then how it reduces the computational time in processing the text data thereby saving time. The main aim of this project is to address the problem of identifying sarcasm in messages and posts in social media dataset, by taking limited quantity of text data (100-200 characters). This project comes up with a methodology to identify sarcasm using emojis.

II. PROBLEM STATEMENT

Sarcasm is "a sharp, bitter, or cutting expression or remark; a bitter gibe or taunt" says, Wikipedia. It can also be defined as the use of irony to mock or convey contempt. It is generally used to refer to the opposite of what you mean in order to make fun of somebody or something or being unpleasant.

Definition: ***Sarcasm Detection Using Emojis***. Comments data from Facebook is scraped and the data containing text + emoji alone is taken, with categorizing the emojis into different types namely, happy, excited, sad, disappointed, angry and sarcastic for a post P. This data is used to train the model and given a new test data (comment) it automatically predicts if the comment is sarcastic or not.

Our approach is to detect sarcasm in comments using the emoji that a user has used along with the text. A text gets life only when it is expressed along with an emoji. The emoji has a better effect in conveying the mental state of a user along with their tone of speech. The percentage of usage of emojis in social media has increased at a high rate and people are able to respond to a post more easily by using emojis, expressing their feelings. This model is built to analyse the effect of emoji in classifying the comments as sarcastic or not along with other additional features.

III. PROPOSED METHOD

The data is obtained by scraping the sarcastic pages from social media websites like Twitter and Facebook. The data containing only the text plus emoji is extracted. The data is preprocessed, and sentiment scores are assigned for the text. The different emojis used in the text are identified and are classified into positive and negative and neutral emojis. The positive emojis are sub divided into happed, excited and negative are sad or disappointed, angry. The features for determining the text sentiment are used to train the classifier models and the performance is reported. The emoji feature is added in addition to the text sentiment features and classifier models are run and the performance of the models are reported. The proposed method is given as a Fig.1 below.

IV. LIST OF TASKS

A. Data Collection

Scraping data from Facebook and Twitter (Jayashree).

B. Data segregation:

Separating data into 3 categories (Varun)

i. Emoji + Text

ii. Emoji only

iii. Text only

C. Data preprocessing

Removing hyperlinks and choosing text data with length less than 200. (Jayashree, Varun)

D. Sentiment Scoring for text and Emoji Classification

Using Emoji + Text data to assign sentiment scores for text and emoji respectively. Sentiment scores are assigned using Regular Expressions in python and Emojis are classified into four categories namely, positive, negative, neutral and the rest as unclassified using Python. Furthermore, positive and negative are categorized as Happy, Excited, Sad, Disappointed, Angry. (Jayashree)

E. Feature Extraction and Feature Selection

Important features for detecting sarcasm (text sentiment score, emoji score) for the comments are selected. Other features like the adjectives used, letters in CAPS, bigrams, trigrams, quadgrams are extracted for the given text. (Varun)

F. Classification

Classification algorithms like SVM, Naive Bayes classifier, Decision Tree classifier, Logistic Regression, Stochastic Gradient Descent Classifier, Random Forest, Adaboost, Gradient Boosting Classifier, Extra Trees Classifier, K-neighbours Classifier are used to test the accuracy for the dataset and the results are reported. (Jayashree, Varun)

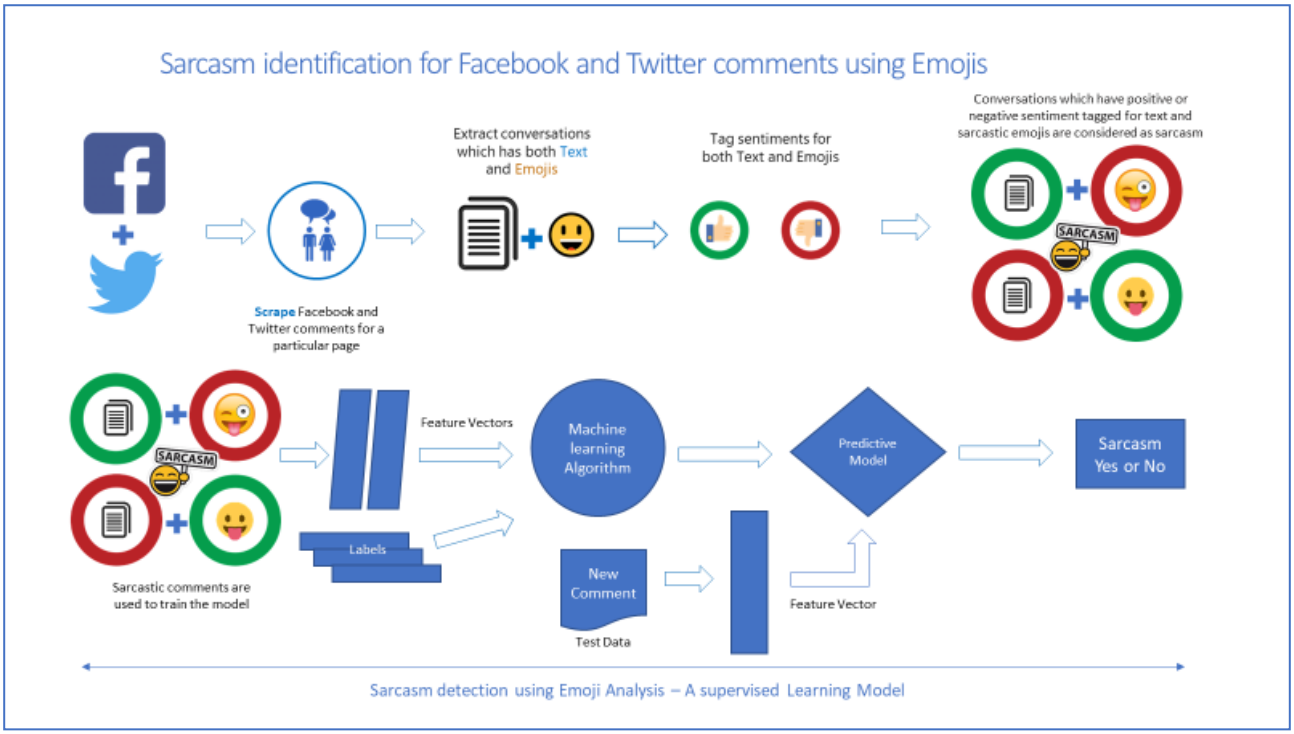

Fig 1. Proposed Method

V. EXPERIMENTS

In this section we did a series of experiments. First was the data collection process. Extracted the text plus emoji data from sarcastic pages in Facebook and Twitter. Obtained the features and determined the contribution of different features to the performance of the model. Compared how the model performed in detecting sarcasm with text alone with how it performed on introducing an emoji feature to it. Finally, we reported and tabulated the results obtained by two methods for a set of classifier algorithms.

Data Collection:

Data was collected from Twitter and Facebook using web scraping in Python. The sarcastic pages like 'sarcasmLOL', 'sarcasmBro' from Facebook and Twitter were chosen and data was scraped, preprocessed and extracted only the data that contains text plus emoji. The data preprocessing involved removal of hyperlinks, special characters, hashtags, retweets, etc.

Feature Extraction:

Features were extracted using the sentiment scoring for the text. These text related features were used to train the classifier models. These features including the different levels of positive and negative keywords used, adjectives describing the qualities, different levels of positive and negative CAPS words, and different combination of these were used to train the model and was tested against the test data. The emoji feature was then introduced along with the text and the model was trained using supervised learning algorithms.

Performance Evaluation:

The model was trained with text features alone for a set of algorithms and compared against the results obtained on adding the emoji as a feature. Performed k-fold cross validation on the datasets for the different classifier algorithms and the F1 scores were obtained.

Classifier algorithms like SVM, Decision Tree classifier, Logistic Regression, Stochastic Gradient, Extra trees classifier, K neighbors classifier, Random forest classifier were run, and the results were tabulated.

VI. ANALYTICAL RESULTS

A. *Examples of sarcastic comments with emojis*

Hahahaha your intelligence 😜😜😜

U can do this, I trust you 😜

Wow she got PHD on selfie taking🍆🍆

Awesome 😜😜

Its big proud 😜😜

Nice post. Will try it 😝

best joke ever😂😜

This talent I don't have.. pls teach me Prani😜😉

Sunny Yadav wow wat an assumption😝

Aqsa Naveed we were proud backbenchers 😜😜😂

Hi guys. I'm so happy and proud of myself and I thought I should share this with you!!! Today, I saw myself on TV, when I turned it off. 😜😝

The above comments without the emoji convey us a different meaning and are taken in the positive sense, however when used with sarcastic emojis it helps us to find

the tone of speech, the mood of the user and identify sarcasm in a better way.

B. The top 20 emojis used along with their count is given below:

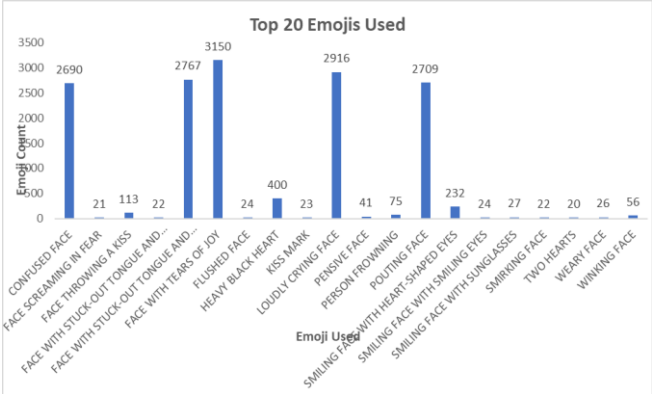

Fig 2. The top 20 emojis used along with their count

It is observed that the Face with stuck-out tongue, Face with tears of Joy and loudly crying face are the most used emojis. This conveys that predominantly the sarcastic comments use these emojis.

C. The sentiment scores for the text are assigend and their counts for the dataset is given below:

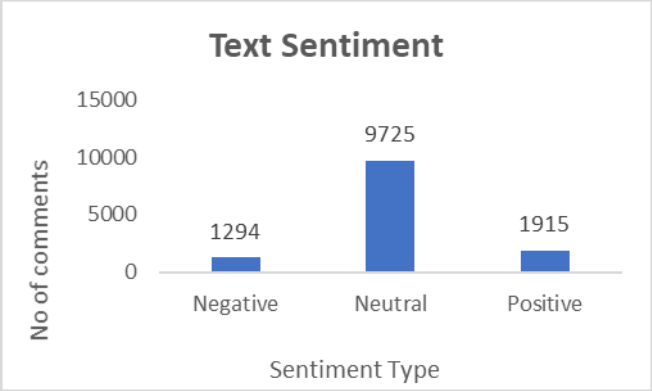

Fig 3. The sentiment scores for the text and their counts

It is observed that the most of comments with text alone falls in neutral category followed by positive.
The k Neighbors Classifier gives the highest F1 score after introducing the emoji feature.

Table 1. Performance evaluation using classification algorithms for text alone

CLASSIFICATION ALGORITHM	F1 SCORE
SVM	0.47733
Decision Tree Classifier	0.47905
Random Forest	0.48568
Adaboost classifier	0.48480
GradientBoostingClassifier	0.48438
K Neighbors Classifier	0.49845
Stochastic Gradient Descent	0.44207
Bayesian Classifier	0.44508
ExtraTreesClassifier	0.48647

Table 2. Performance evaluation using classification algorithms after emoji introduction

CLASSIFICATION ALGORITHM	F1 SCORE
SVM	0.95644
Decision Tree Classifier	0.95777
Random Forest	0.95836
Adaboost classifier	0.95628
GradientBoostingClassifier	0.95734
K Neighbors Classifier	0.96790
Stochastic Gradient Descent	0.95852
Bayesian Classifier	0.92049
ExtraTreesClassifier	0.95783

VII. CONCLUSION AND FUTURE WORK

It is observed that the model performs better and gives better results on introducing the emoji as a feature in addition to text for detecting sarcasm in comments. The model gives around 40% accuracy for text data alone, giving a room for introducing an emoji and introduction of this emoji increases the performance by around 45%. Face with stuck-out tongue, Face with tears of Joy and loudly crying face are the most used emojis. This conveys that predominantly the sarcastic comments use these emojis. It is also observed that the most of comments with text alone falls in neutral category followed by positive. The comments with a positive or negative text sentiment and a sarcastic emoji were found to be sarcastic.

Scope for additional features to train the model. The model can be trained in future for identifying sarcasm to reduce computational time using emojis alone. This analysis can be extended by the usage of CNN, embedding concepts, and other deep learning algorithms.

REFERENCES

[1] Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. Sarcasm Detection on Twitter: A Behavioral Modeling Approach, 2015.

[2] R. Zafarani and H. Liu. Connecting users across social media sites: a behavioral-modeling approach. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 41–49. ACM, 2013.

[3] Dyllan Furness. MIT algorithm learns to understand emotions and sarcasm throughemojis, 2017.

[4] Sharavan. Sentiment Analysis in Python using NLTK, 2016.

[5] H. P. Grice. Some further notes on logic and conversation. In P. Cole, editor, Syntax and Semantics 9: Pragmatics, pages 113–127. 1978.

[6] O. Tsur, D. Davidov, and A. Rappoport. Icwsm-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In ICWSM, 2010.