CKME136 Capstone Project Final Report

Project Title : Twitter Text Sentiment Analysis

Author : Hiren Dossani

Date : July 30, 2019

Professors : Dr. Ceni Babaoblu, Dr. Tamer Abdou

Project Repository : <https://github.com/hdossani/ckme136_capstone_project>

# Introduction

Twitter is a very popular social networking and microblogging

service where people communicate in short messages called tweets. This simple, but effective service has

- insights from 83% of world leaders

- 330 million monthly users

- 3 billion account holders

It’s no wonder so many companies want to tap into the potential of Twitter for their own social strategies. The fast-paced nature of this platform means that it’s a great way for brands to start building a stronger online presence. Learning how to track the right twitter metrics could help you make insightful decisions about your future marketing campaigns. The more you learn about your audience, the more you can transform every interaction into an actionable strategy for success.

Sentiment Analysis is a field within Natural Language Processing (NLP) that builds systems that try to identify and extract opinion within text. Usually, besides identifying the opinion, these systems extract attributes of the expression e.g.

- Polarity : positive, negative or neutral opinion

- Subject : Entity that is talked about

- Opinion Holder : Person or an entity that expresses the opinion

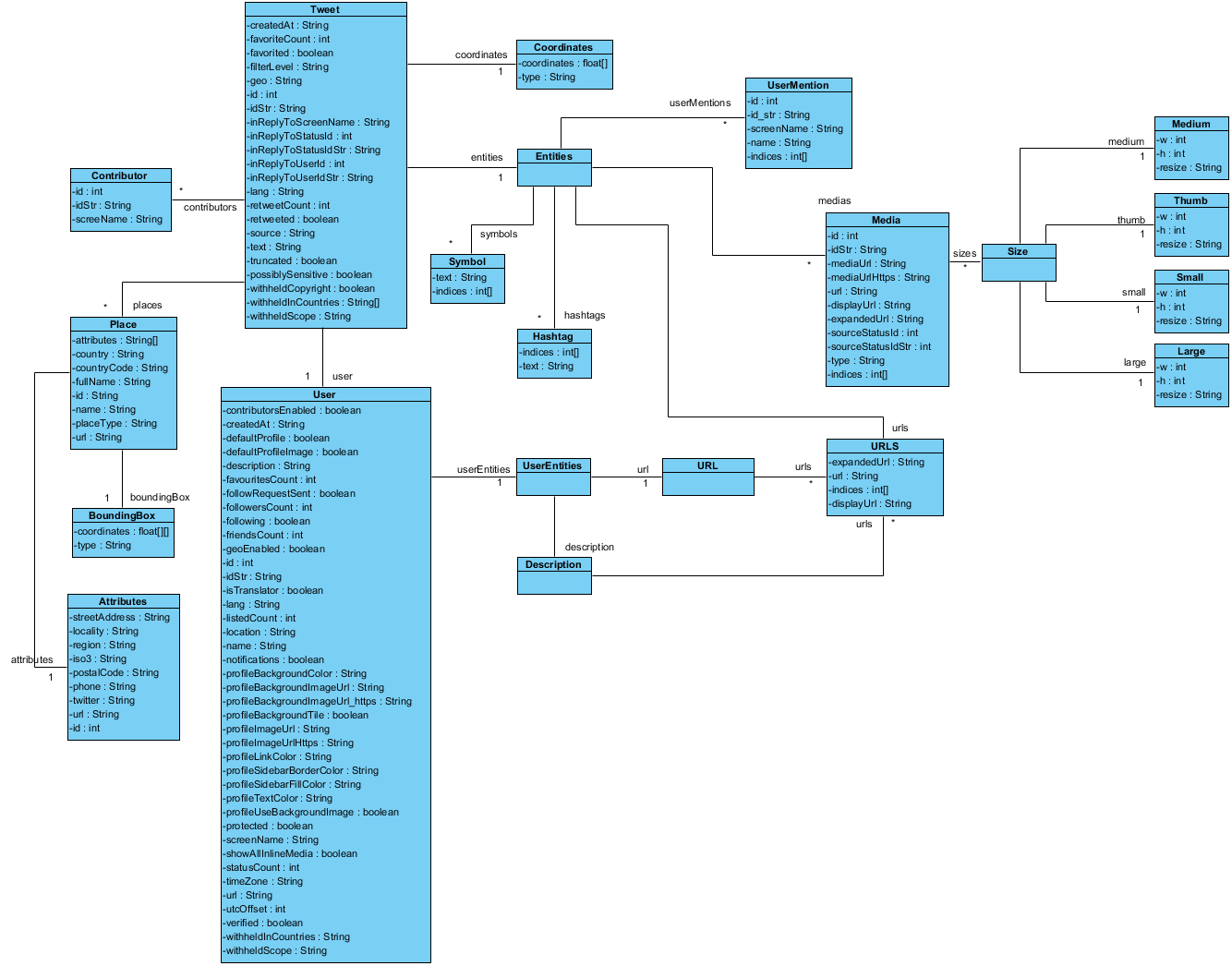
Twitter Sentiment Analysis can be used to identify the customer or follower's attitude towards a brand through the use of variables such as context, tone, emotion, etc. Marketers can use sentiment analysis to research public opinion of their company and products, or to analyze customer satisfaction. Organizations can also use this analysis to gather critical feedback about problems in newly released products. Sentiment analysis not only helps companies understand how they’re doing with their customers, it also gives them a better picture of how they stack up against their competitors. Knowing the sentiments associated with competitors helps companies evaluate their own performance and search for ways to improve. Even political parties and politicians use twitter sentiment analysis to picture how they stack up against their competitors.

The aim of this project is to develop a functional classifier for accurate and automatic sentiment classification of an unbounded tweet stream using AWS Cloud Services

# Understanding Twitter Data

Dataset for this project is an unbounded stream of live tweets.

The data structure of each tweet is as below.



The above picture displays all the elements, relationships between elements and element datatype of a single tweet. A single tweet looks like:

{'created\_at': 'Sun Jul 28 22:36:05 +0000 2019', 'id': 1155607889656107008, 'id\_str': '1155607889656107008', 'text': 'RT @remvcc: This is how Justin Trudeau reduces his plastic waste https://t.co/DBzx57wwa1 via @YouTube', 'source': '<a href="http://twitter.com/#!/download/ipad" rel="nofollow">Twitter for iPad</a>', 'truncated': False, 'in\_reply\_to\_status\_id': None, 'in\_reply\_to\_status\_id\_str': None, 'in\_reply\_to\_user\_id': None, 'in\_reply\_to\_user\_id\_str': None, 'in\_reply\_to\_screen\_name': None, 'user': {'id': 1686575868, 'id\_str': '1686575868', 'name': 'Benny', 'screen\_name': '4ff2', 'location': None, 'url': None, 'description': None, 'translator\_type': 'none', 'protected': False, 'verified': False, 'followers\_count': 159, 'friends\_count': 186, 'listed\_count': 7, 'favourites\_count': 11004, 'statuses\_count': 11229, 'created\_at': 'Tue Aug 20 20:10:11 +0000 2013', 'utc\_offset': None, 'time\_zone': None, 'geo\_enabled': False, 'lang': None, 'contributors\_enabled': False, 'is\_translator': False, 'profile\_background\_color': 'C0DEED', 'profile\_background\_image\_url': 'http://abs.twimg.com/images/themes/theme1/bg.png', 'profile\_background\_image\_url\_https': 'https://abs.twimg.com/images/themes/theme1/bg.png', 'profile\_background\_tile': False, 'profile\_link\_color': '1DA1F2', 'profile\_sidebar\_border\_color': 'C0DEED', 'profile\_sidebar\_fill\_color': 'DDEEF6', 'profile\_text\_color': '333333', 'profile\_use\_background\_image': True, 'profile\_image\_url': 'http://pbs.twimg.com/profile\_images/1026862286139314176/gZvNoNP1\_normal.jpg', 'profile\_image\_url\_https': 'https://pbs.twimg.com/profile\_images/1026862286139314176/gZvNoNP1\_normal.jpg', 'default\_profile': True, 'default\_profile\_image': False, 'following': None, 'follow\_request\_sent': None, 'notifications': None}, 'geo': None, 'coordinates': None, 'place': None, 'contributors': None, 'retweeted\_status': {'created\_at': 'Sun Jul 28 22:33:53 +0000 2019', 'id': 1155607335160516608, 'id\_str': '1155607335160516608', 'text': 'This is how Justin Trudeau reduces his plastic waste https://t.co/DBzx57wwa1 via @YouTube', 'source': '<a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>', 'truncated': False, 'in\_reply\_to\_status\_id': None, 'in\_reply\_to\_status\_id\_str': None, 'in\_reply\_to\_user\_id': None, 'in\_reply\_to\_user\_id\_str': None, 'in\_reply\_to\_screen\_name': None, 'user': {'id': 897157166, 'id\_str': '897157166', 'name': 'REM - TRUDEAU OUT ! - NO U.N. COMPACT !', 'screen\_name': 'remvcc', 'location': 'French parents, born in Mexico, proud Canadian by choice. Conservative. Not PC. Pro Free-Speech. Pro Israel Zionist. Anti-Trudeau Pro-Trump.', 'url': None, 'description': '🇨🇦🇲🇽🇮🇱🇫🇷🇺🇸', 'translator\_type': 'none', 'protected': False, 'verified': False, 'followers\_count': 2408, 'friends\_count': 4946, 'listed\_count': 1, 'favourites\_count': 16517, 'statuses\_count': 16391, 'created\_at': 'Mon Oct 22 09:29:36 +0000 2012', 'utc\_offset': None, 'time\_zone': None, 'geo\_enabled': True, 'lang': None, 'contributors\_enabled': False, 'is\_translator': False, 'profile\_background\_color': 'C0DEED', 'profile\_background\_image\_url': 'http://abs.twimg.com/images/themes/theme1/bg.png', 'profile\_background\_image\_url\_https': 'https://abs.twimg.com/images/themes/theme1/bg.png', 'profile\_background\_tile': False, 'profile\_link\_color': '1DA1F2', 'profile\_sidebar\_border\_color': 'C0DEED', 'profile\_sidebar\_fill\_color': 'DDEEF6', 'profile\_text\_color': '333333', 'profile\_use\_background\_image': True, 'profile\_image\_url': 'http://pbs.twimg.com/profile\_images/1072698119618408448/MU1zVW0L\_normal.jpg', 'profile\_image\_url\_https': 'https://pbs.twimg.com/profile\_images/1072698119618408448/MU1zVW0L\_normal.jpg', 'profile\_banner\_url': 'https://pbs.twimg.com/profile\_banners/897157166/1548183314', 'default\_profile': True, 'default\_profile\_image': False, 'following': None, 'follow\_request\_sent': None, 'notifications': None}, 'geo': None, 'coordinates': None, 'place': None, 'contributors': None, 'is\_quote\_status': False, 'quote\_count': 0, 'reply\_count': 0, 'retweet\_count': 1, 'favorite\_count': 0, 'entities': {'hashtags': [], 'urls': [{'url': 'https://t.co/DBzx57wwa1', 'expanded\_url': 'https://youtu.be/LMJ8UJ8wEiY', 'display\_url': 'youtu.be/LMJ8UJ8wEiY', 'indices': [53, 76]}], 'user\_mentions': [{'screen\_name': 'YouTube', 'name': 'YouTube', 'id': 10228272, 'id\_str': '10228272', 'indices': [81, 89]}], 'symbols': []}, 'favorited': False, 'retweeted': False, 'possibly\_sensitive': False, 'filter\_level': 'low', 'lang': 'en'}, 'is\_quote\_status': False, 'quote\_count': 0, 'reply\_count': 0, 'retweet\_count': 0, 'favorite\_count': 0, 'entities': {'hashtags': [], 'urls': [{'url': 'https://t.co/DBzx57wwa1', 'expanded\_url': 'https://youtu.be/LMJ8UJ8wEiY', 'display\_url': 'youtu.be/LMJ8UJ8wEiY', 'indices': [65, 88]}], 'user\_mentions': [{'screen\_name': 'remvcc', 'name': 'REM - TRUDEAU OUT ! - NO U.N. COMPACT !', 'id': 897157166, 'id\_str': '897157166', 'indices': [3, 10]}, {'screen\_name': 'YouTube', 'name': 'YouTube', 'id': 10228272, 'id\_str': '10228272', 'indices': [93, 101]}], 'symbols': []}, 'favorited': False, 'retweeted': False, 'possibly\_sensitive': True, 'filter\_level': 'low', 'lang': 'en', 'timestamp\_ms': ‘1564353365707'}

The tweet is in JSON format and for text sentiment analysis, we are only interested in tweet text.

# Project Architecture

Project architecture uses AWS managed services. Services used in this project are Kinesis, S3, Comprehend, Athena and Quicksight to build a Natural Language Processing (NLP) powered solution. It uses serverless data processing and machine learning (ML) pipeline that processes stream of tweets for sentiment analysis of a given entity.

e.g. Liberals political party or Trudeau, etc.

Here’s a brief description of each AWS services used:

* Kinesis : Easily capture, prepare, and load real-time data streams into data stores e.g. S3
* S3 : Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance.
* Comprehend : Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to find insights and relationships in text.
* Athena : Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage
* QuickSight : Fast, cloud-powered business intelligence service that makes it easy to deliver insights to everyone

Stream of Tweets

Kinesis Firehose

S3 Raw Data

Lambda Function:

Clean Data

Comprehend Detect Sentiment

Comprehend Detect Entity

S3 Sentiment

S3 Entity

Athena

QuickSight

# Data Cleaning

The preprocessing of the text data is an essential step as it makes the raw text ready for mining, i.e., it becomes easier to extract information from the text and apply machine learning algorithms to it. If we skip this step then there is a higher chance that you are working with noisy and inconsistent data. The objective of this step is to clean noise those are less relevant to find the sentiment of tweets such as special characters, usernames and URLs which don’t carry much weightage in context to the text.

Example of cleaned tweet text :

{"tweetid": 1155683195289030657, "text": " Prime Minister Justin Trudeau has turned Canada into an economic superstar!\nJuly 12th, 2019\nScotiabank of Can”}

Cleaned data is put into AWS Kinesis Firehose. Firehose, then saves the data in S3 bucket named ckme136.capstone.twitter

# AWS Lambda

AWS Lambda lets you run code without provisioning or managing servers. Lambda takes care of everything required to run and scale code with high availability. You can set up code to automatically trigger from other AWS services or call it directly from any web or mobile app. In our case, lambda code is triggered by saving data in S3 bucket ckme136.capstone.twitter.

Lambda code invokes AWS Comprehend service to detect sentiment and entity for every tweet record it reads from S3 bucket ckme136.capstone.twitter.

Sentiment analysis and entities detected are put into AWS Kinesis Firehose. Firehose saves the data in S3 buckets namely ckme136.capstone.twitter.sentiment and ckme136.capstone.twitter.entity respectively.

# AWS Comprehend

Amazon Comprehend uses natural language processing (NLP) to extract insights about the content of documents. Amazon Comprehend processes any text file in UTF-8 format. It develops insights by recognizing the entities, key phrases, language, sentiments, and other common elements in a document.

Amazon Comprehend uses a pre-trained model to examine and analyze a document or set of documents to gather insights about it. This model is continuously trained on a large body of text so that there is no need for you to provide training data. Amazon takes care of the infrastructure to host vast amount of data, data changes, context changes, new types of data being added, etc. In nutshell the feedback loop is continuous.

# Training & Prediction

The process of training an ML model involves providing an ML algorithm (that is, the *learning algorithm*) with training data to learn from. The term *ML model* refers to the model artifact that is created by the training process.

The training data must contain the correct answer, which is known as a *target* or *target attribute*. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict), and it outputs an ML model that captures these patterns.

You can use the ML model to get predictions on new data for which you do not know the target.

With AWS Comprehend, no need to provide training data.

# Sentiment Analysis Algorithms

Classification is a stage in sentiment analysis that can described as a process in which we predict qualitative response, or in this case we classify the document into its polarity. Predicting a qualitative response of a document can be referred to as classifying the document since it involves assigning an observation to a category or class. There are many possible classification techniques, or classifiers that one might use for to predict the qualitative response or class of a document. In sentiment analysis some widely used classification techniques are as follows:

ML models for multiclass classification problems allow you to generate predictions for multiple classes (predict one of more than two outcomes). For training multiclass models, Amazon ML uses the industry-standard learning algorithm known as multinomial logistic regression.

# Accuracy of the Result

Correct Predictions

* True positive (TP): predicted the value as 1, and the true value is 1.
* True negative (TN): predicted the value as 0, and the true value is 0.

Erroneous Predictions

* False positive (FP): predicted the value as 1, but the true value is 0.
* False negative (FN): predicted the value as 0, but the true value is 1.
* Accuracy

The fraction of the labels that were correct recognized. It is computed by dividing the number of labels in the test documents that were correctly recognized by the total number of labels in the test documents.



* F1Score

A measure of how accurate the classifier results are for the test data. It is derived from the Precision and Recall values. The F1Score is the harmonic average of the two scores. The highest score is 1, and the worst score is 0.

* Precision

A measure of the usefulness of the classifier results in the test data. High precision means that the classifier returned substantially more relevant results than irrelevant ones.

* Recall

A measure of how complete the classifier results are for the test data. High recall means that the classifier returned most of the relevant results.



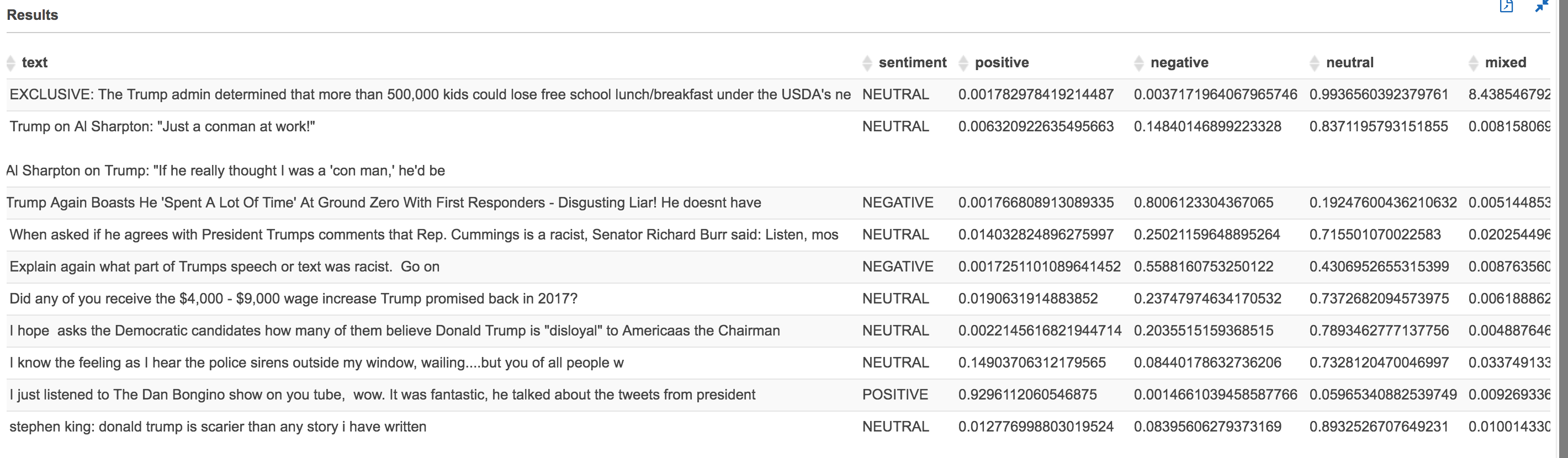
# Challenges of Sentiment Analysis

The key challenges for sentiment analysis are:-

* Named Entity Recognition - What is the person actually talking about, e.g. is 300 Spartans a group of Greeks or a movie?
* Anaphora Resolution - the problem of resolving what a pronoun, or a noun phrase refers to. "We watched the movie and went to dinner; it was awful." What does "It" refer to?
* Parsing - What is the subject and object of the sentence, which one does the verb and/or adjective actually refer to?
* Sarcasm - If you don't know the author you have no idea whether 'bad' means bad or good.
* Twitter - abbreviations, lack of capitals, poor spelling, poor punctuation, poor grammar, ...

# Results of Sentiment Analysis on Twitter Stream of 60 secs on Trump!

# 



# References

|  |  |
| --- | --- |
| AWS Services | <https://docs.aws.amazon.com/index.html> Amazon managed services such as S3, Athena, Lambda, Kinesis, Comprehend, QuickSight to perform language natural language processing on the tweets flowing through the system and to create data visualization. |
| Like It or Not: A Survey of Twitter Sentiment Analysis Methods | <https://dl.acm.org/citation.cfm?id=2938640> This survey provides an overview of the topic by investigating and briefly describing the algorithms that have been proposed for sentiment analysis in Twitter |
| A real-time Twitter sentiment analysis | (<https://dl.acm.org/citation.cfm?id=3102282> A real-time implementation of a system that can discover and track opinions on Twitter. A recommended approach using machine learning techniques, to analyze opinions and detect tweets polarity. |
| Python For Everybody | [www.py4e.com/html3](http://www.py4e.com/html3) |
| AWS Machine Learning | <https://docs.aws.amazon.com/machine-learning/latest/dg/what-is-amazon-machine-learning.html> |