

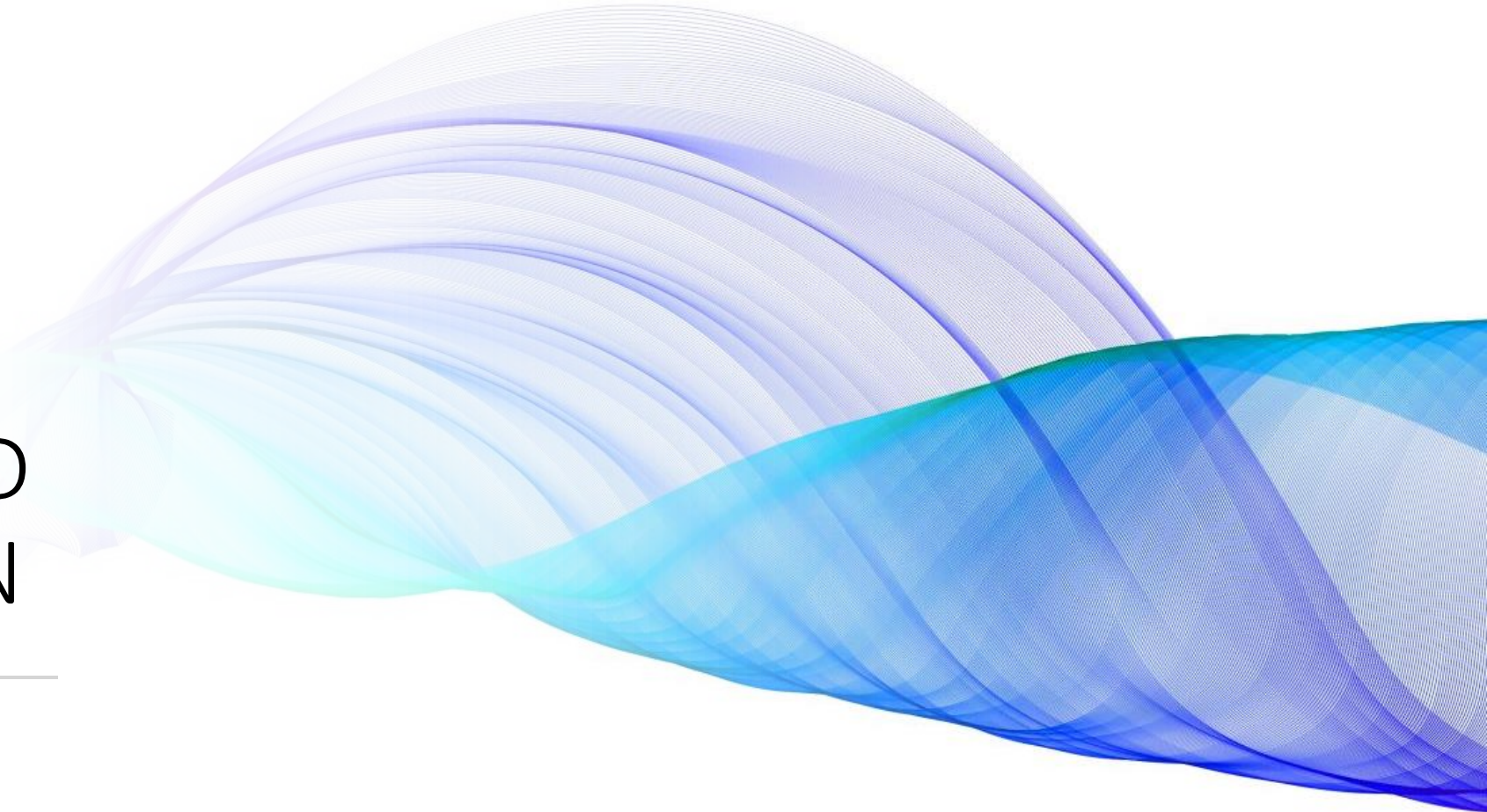


# TWEET EXTRACTION AND SUMMARIZATION

---

Group 1

Chaitanya Agarwal, Tanvi Kamble,  
Apurvi Mansinghka, Jyoti Gambhir





# Problem Statement

Given a set of hashtags, scrape Twitter for tweets related to them and analyze (summarization, sentiment analysis etc.) the scraped content.



# Brief Introduction

- TwitterScraper for tweet extraction
- Dataset including 11K pre-processed tweets centered on the 'black lives matter' movement
- Text rank algorithm for summarization
- Sentiment Analysis on extracted tweets using Huggingface BERT Model and other baseline models
- The results obtained are satisfactory and we also take care of the problems faced



# Scraping Tweets

- Initial experiments with Tweepy
  - Retrieve tweets upto 7 days
  - Scraping 18,000 tweets per a 15 min window
- TwitterScraper
  - Extraction is limited by internet speed/bandwidth
  - Threaded REST Calls
  - Popular tweets based on the following two hashtags were scraped:  
*#blacklivesmatter, #blm*
  - 22K tweets obtained



# Preprocessing

- Removal of duplicate tweets
- Replacement of abbreviations with their expansions
- Link and mentions removal
- Contraction replacement
- Hashtag removal
- Twitter meta including '*RT*', and username removal
- Punctuation and numeral handling
- Persisting the dataset as a Pandas Dataframe

# Baselines

Work	Dataset (size of each cluster)	System Output	Evaluation Metrics
Inouye and Kalita (2011)	trending topics (approximately 1500 tweets)	4 tweets	ROUGE and human (overall quality comparing to human summary)
Sharifi et al. (2010)	same as above	1 tweet	same as above
Rosa et al. (2011)	segmented hashtag topics by LDA and k-means clustering (average 410 tweets)	1, 5, 10 tweets	Precision@k (relevance to topic)
Harabagiu and Hickl (2011)	real-word event topics (a minimum of 2500 tweets)	top tweets until a limit of 250 words was reached	human (coverage and coherence)
Liu et al. (2011a)	general topics and hashtag topics (average 1.7k tweets)	same lengths as of the human summary, vary for each topic (about 2 or 3 tweets)	ROUGE and human (content coverage, grammaticality, non-redundancy, referential clarity, focus)
Wei et al. (2012)	segmented hashtag topics according to burstiness (average 10k tweets)	10 tweets	ROUGE, Precision/Recall (good readability and rich content)
Takamura et al. (2011)	specific soccer games (2.8k - 5.2k tweets)	same lengths as the human summary, vary for each topic (26 - 41 tweets)	ROUGE (considering only content words)
Chakrabarti and Punera (2011)	specific football games (1.8k tweets)	10 - 70 tweets	Precision@k (relevance to topic)

## A Preliminary Study of Tweet Summarization using Information Extraction

**Wei Xu, Ralph Grishman, Adam Meyers**  
 Computer Science Department  
 New York University  
 New York, NY 10003, USA  
 {xuwei,grishman,meyers}@cs.nyu.edu

**Alan Ritter**  
 Computer Science and Engineering  
 University of Washington  
 Seattle, WA 98125, USA  
 aritter@cs.washington.edu

### Abstract

Although the ideal length of summaries differs greatly from topic to topic on Twitter, previous work has only generated summaries of a pre-fixed length. In this paper, we propose an event-graph based method using information extraction techniques that is able to create summaries of variable length for different topics. In particular, we extend the Pagerank-like ranking algorithm from previous work to partition event graphs and thereby detect fine-grained aspects of the event to be summarized. Our preliminary results show that summaries created by our method are more concise and news-worthy than SumBasic according to human judges. We also provide a brief survey of datasets and evaluation design used in previous work to highlight the need of developing a standard evaluation for automatic tweet summarization task.

### 1 Introduction

Tweets contain a wide variety of useful information from many perspectives about important events taking place in the world. The huge number of messages, many containing irrelevant and redundant information, quickly leads to a situation of information overload. This motivates the need for automatic summarization systems which can select a few messages for presentation to a user which cover the most important information relating to the event without redundancy and filter out irrelevant and personal information that is not of interest beyond the user's immediate social network.

Although there is much recent work focusing on the task of multi-tweet summarization (Becker et al., 2011; Inouye and Kalita, 2011; Zubiaga et al., 2012; Liu et al., 2011a; Takamura et al., 2011; Harabagiu and Hickl, 2011; Wei et al., 2012), most previous work relies only on surface lexical clues, redundancy and social network specific signals (e.g. user relationship), and little work has considered taking limited advantage of information extraction techniques (Harabagiu and Hickl, 2011) in generative models. Because of the noise and redundancy in social media posts, the performance of off-the-shelf news-trained natural language process systems is degraded while simple term frequency is proven powerful for summarizing tweets (Inouye and Kalita, 2011). A natural and interesting research question is whether it is beneficial to extract named entities and events in the tweets as has been shown for classic multi-document summarization (Li et al., 2006). Recent progress on building NLP tools for Twitter (Ritter et al., 2011; Gimpel et al., 2011; Liu et al., 2011b; Ritter et al., 2012; Liu et al., 2012) makes it possible to investigate an approach to summarizing Twitter events which is based on Information Extraction techniques.

We investigate a graph-based approach which leverages named entities, event phrases and their connections across tweets. A similar idea has been studied by Li et al. (2006) to rank the salience of event concepts in summarizing news articles. However, the extreme redundancy and simplicity of tweets allows us to explicitly split the event graph into subcomponents that cover various aspects of the initial event to be summarized to create comprehen-



# Dataset

Number of Tweets 11,205

Duplicate Tweets : 11,077

Total Words : 1,81,835

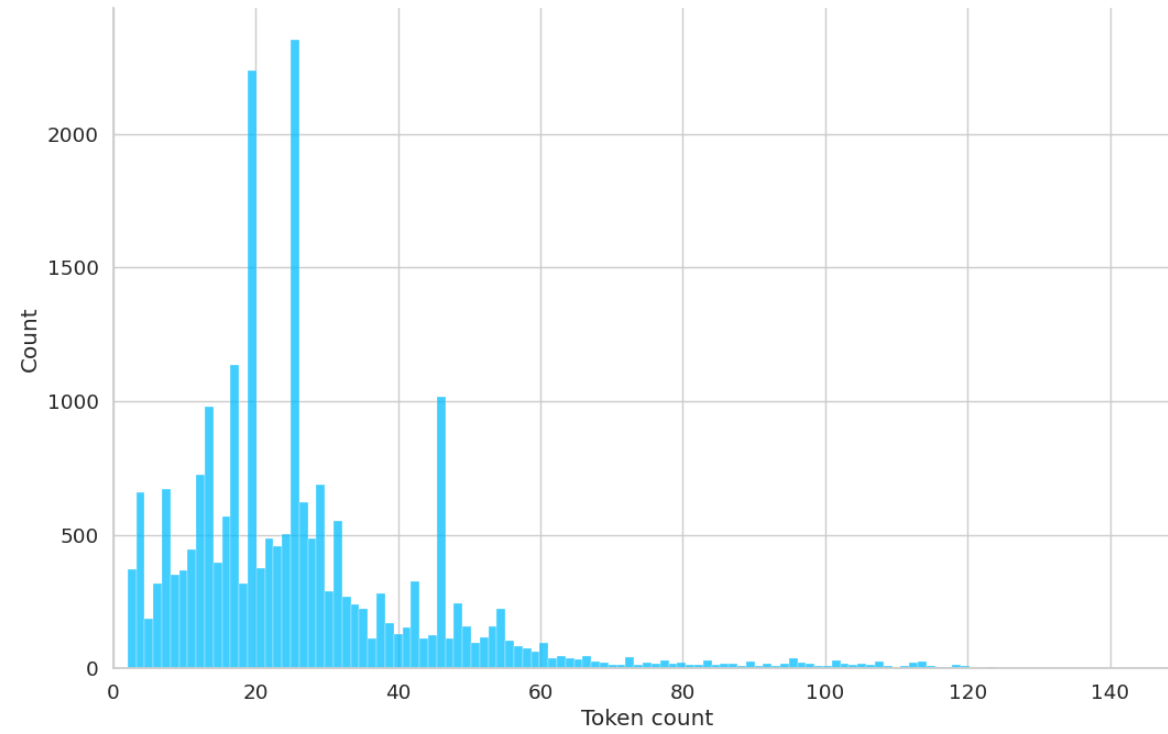
Longest Tweet: 60 Words

Unique Words: 26,190

Hashtags: 5,707

Mentions: 23,811

Punctuation Marks: 2,049



Length of  
Tweets





# Text Rank

- Unsupervised Algorithm to rank given sentences based on a graph-based representation of word vectors
- Utilized 100D Pretrained GloVe Embeddings trained on 4B tokens
- Unlike the baselines which output  $n$  most contextual tweets, it outputs  $n$  most contextual sentences
- Results for  $n=3$ 
  - You hear them fighting for BLACK LIVES MATTER but enforcement still makes no step. No justice, no peace, take it to the streets and f\*\*\* the police. Protests erupt in Philadelphia after police shoot armed black man.



# Sentiment Analysis

- Finding the Sentiments of the extracted tweets Using TextBlob

Sentiment	Percentage of Tweets
Positive	14.3
Negative	20.2
Neutral	65.5

# Tweets Classified by the TextBlob

## Positive tweets:

Did anyone notice the antifa BLM coverage and antics has been sparse to none since the first debate  
BLM murals one 1200 check some racist statues down some snarky tweets and being told to wash our hands sure has given  
The police scanner s pretty much a blank right now in Philadelphia Police have switched to tactical communications because  
PHILADELPHIA Plagued by years of a far left leadership These same thugs watch while Phillys businesses burn WHY  
In terms of her art style Trump biden harris equality blm dump trump shirt Aya Kato says she intends to put the world of harmony and give it a new sense  
17 KAMU YG SELALU GERCEP MEMBALAS ORG YG BUTUH BANTUAN Pemilik kost ternyaman yg dijadiin basecamp makan2 Gatau kenapa suka aja gtu kalo curhat sama ni anak Diem2 trnyata banyak yg sepik t  
p skrg dah sold out Haha kasian blm ngedate offline kayanya ya Terajin klo ditanya2 tgs  
New Sheriff Luciano  
Delighted to hear that our proposal to to investigate GDM in women from subsaharan Africa has been funded Really looking forward to working with at on this important question cc  
I do not think you are you have missed the point completely but I sort of already guessed that you would Anyway I have obviously triggered you but I cannot sit here with you monopolisin  
g my time I have things to discuss like the curse of Islam and how soggy BLM are  
PHILADELPHIA Plagued by years of a far left leadership These same thugs watch while Phillys businesses burn WHY

## Negative tweets:

A black man Walter Wallace yielded a knife at police in Philadelphia and was shot and killed Riots broke out by BLM  
Our kids are already experiencing racism fourth grade teacher Melissa Statz said Our Black and brown students are dealing  
Read this article This is still happening You hear them fighting for BLACK LIVES MATTER but enforcement still makes i  
Jared Kushner who has repeatedly criticized BLM protests did it again today and said this about black people and President T  
Yah the Joe Biden campaign is OVER Beware of violent BLM s coming which works s Joe Biden s Campaign or Rally as they have done for months  
Black Lives Matter rioters just hit a cop in Philadelphia with a truck  
Helping the cause So very very sad  
The Fake News media continues to ignore evidence of the Biden Crime Familys corrupt deals with foreign business leaders  
Jared Kushner who has repeatedly criticized BLM protests did it again today and said this about black people and President T  
The British MSM ANTI Trump Boris Government British Majority Democracy History Justice Union PRO EU Illegal immigrat



# Training Models for Sentiment Analysis

- Train the Models on Naïve Bayes and Logistic Regression
- Results obtained:
  - Accuracy in Naive Bayes: **91%**
  - Accuracy in Logistic Regression Model: **92%**



# Using BERT by HuggingFace

- First, we use BertTokenizer to create Token IDs
- Training on BERT and Transformer given by HuggingFace
- After Training: (10th Epoch)
  - Train loss 0.00065  
Accuracy 0.99
  - Val loss 0.29502  
Accuracy 0.966