Term Projects: Basic Requirements

- Your term project must be an **NLP application**, dealing primarily with **text data**, which could be any one of the existing, publicly available corpora. You may collect your own data as well.
  - (**Optional**) You may combine your text data with non-text data, such as social network data, image/video data, etc., in a creative way, to make your project more interesting.
  - Review my slides, if needed, to refresh your memory about the official definition of NLP and NLP applications. But you should feel free to be creative.
  - Slide 5 shows a sample NLP project by my former TA, Nikhil, for his NLP class. I have also selected 4 NLP-related projects from my Social Media Data Mining class (slides 7-24).
- Form a group of **5** people.

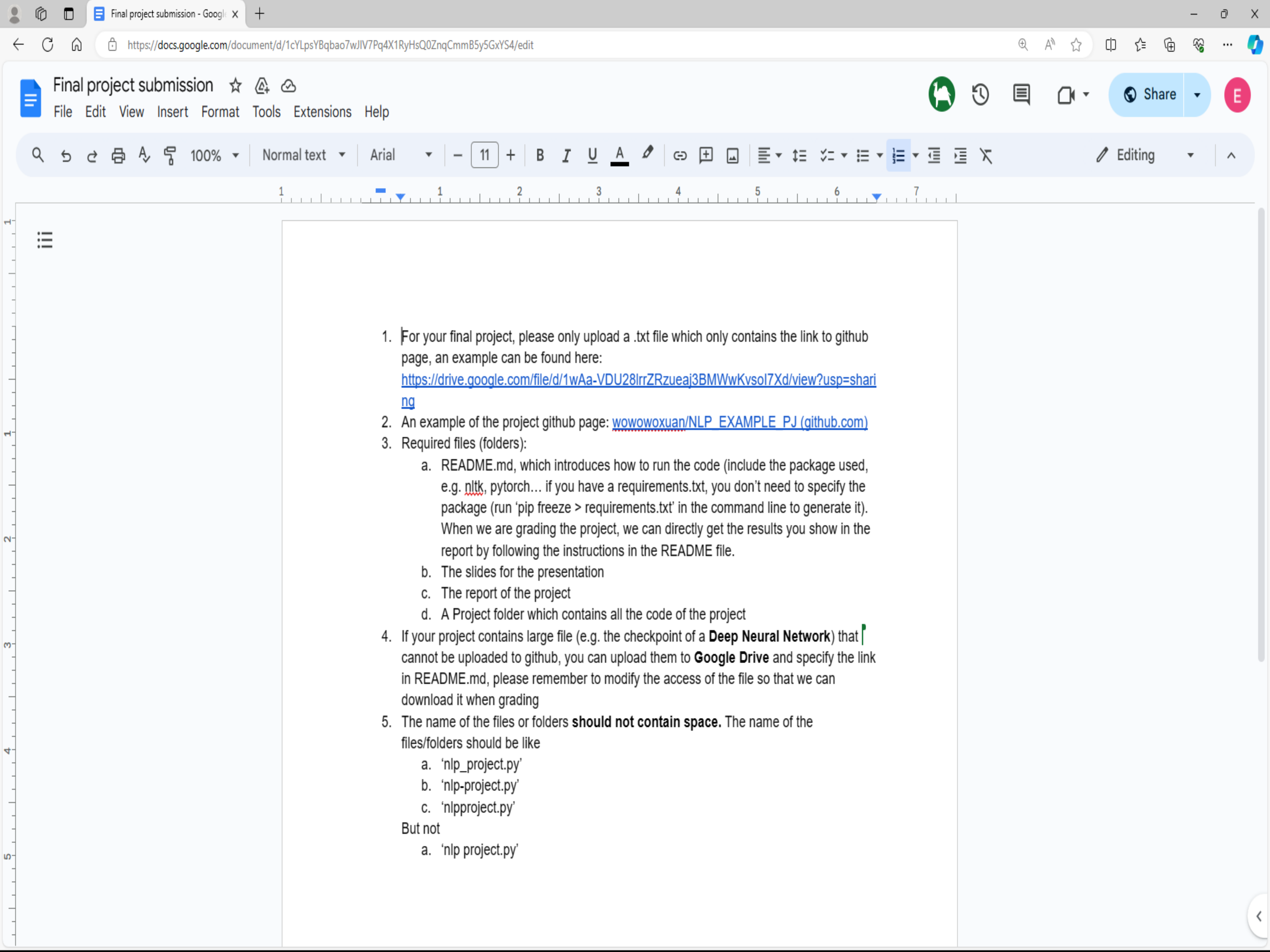
Term Project Proposal & Deliverables

- Term project proposals due **March 31.** At a minimum, you should submit 2 full pages (single-spaced with font size no larger than 12 points) describing:
  1. Your project **idea** and **features**
  2. **Significance** of your idea (Why do you think that's a good idea?)
  3. **Work plan**: technical tasks, their start/end dates, and who is responsible for which task.
- Will be graded mainly by 3 criteria:
  - **Creativity (2.5 points)**: How special/significant is your project idea?
  - **Completeness (3 points)**: How many (interesting) features have you proposed/described in detail?
  - **Clarity (2.5 points)**: Is your proposal, especially your work plan, clear about what you plan to accomplish?

Term Project Proposal & Deliverables

- Term Project Presentations: **April 23, 25** (More details to come)
- Term project deliverables due: **May 3**
- **Note**: TA Weiheng has further refined my rules in the syllabus for submitting your term project deliverables (See next slide). Please following them.
  - Its URL:

https://docs.google.com/document/d/1cYLpsYBqbao7wJlV7Pq4X1RyHsQ0ZnqCmmB5y5GxYS4/edit


Final project submission - Googl
https://docs.google.com/document/d/1cYLpsYBqbao7wJlV7Pq4X1RyHsQ0ZnqCmmB5y5GxYS4/edit
Final project submission
File
Edit
View
Insert
Format
Tools
Extensions
Help
Share
E
100%
Normal text
Arial
11
Editing

1. For your final project, please only upload a .txt file which only contains the link to github page, an example can be found here: https://drive.google.com/file/d/1wAa-VDU28lrrZRzueaj3BMWwKvsol7Xd/view?usp=sharing
2. An example of the project github page: wowowoxuan/NLP_EXAMPLE_PJ (github.com)
3. Required files (folders):
   a. README.md, which introduces how to run the code (include the package used, e.g. nltk, pytorch… if you have a requirements.txt, you don't need to specify the package (run 'pip freeze > requirements.txt' in the command line to generate it). When we are grading the project, we can directly get the results you show in the report by following the instructions in the README file.
   b. The slides for the presentation
   c. The report of the project
   d. A Project folder which contains all the code of the project
4. If your project contains large file (e.g. the checkpoint of a **Deep Neural Network**) that cannot be uploaded to github, you can upload them to **Google Drive** and specify the link in README.md, please remember to modify the access of the file so that we can download it when grading
5. The name of the files or folders **should not contain space.** The name of the files/folders should be like
   a. 'nlp_project.py'
   b. 'nlp-project.py'
   c. 'nlpproject.py'

   But not
   a. 'nlp project.py'

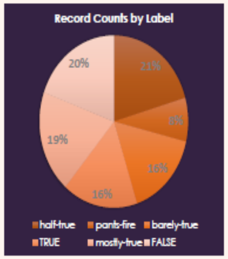
Unmasking Misinformation using NLP Techniques (LIAR Dataset)

Submitted by Nikhil Varma Pratap, Chiradeep Nanabala, IndraneelSomayajula, Shivangi Mundhra

Under Prof. Lu Xiao

Introduction

Fake news is false or misleading information spread through various media channels, causing harm by influencing people's beliefs and actions. It can cause harm by creating confusion, influencing people's beliefs and actions, and spreading mis information. It is a widespread issue with severe consequences, including political polarization and public health misinformation. Detecting fake news using NLP techniques can help prevent its spread and minimize its impact.



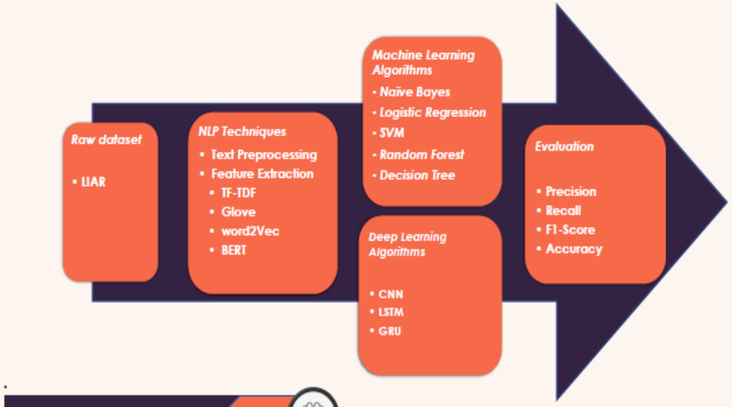
The "LIAR" dataset on Kaggle has 12,836 fact-checked statements labeled with six truthfulness levels. True, Half True, Barely True, Mostly True, Pants on Fire, and False. It helps researchers develop fake news detection algorithms with features like statement, speaker, context, and source. The dataset is balanced with roughly equal numbers of statements in each category. [3]

Methodology

In this project, we will be preprocessing and analyzing data, extracting features(word embeddings) using various methods And, We will be training the model with Machine Learning and Deep Learning models . After prediction, we will be evaluating the models using Accuracy, precision, f1 score thereby selecting the best performing model with best embedding method.

Text Preprocessing is an essential step in building accurate fake news detection models. It involves text cleaning, lowercasing, tokenization, stop word removal, stemming/lemmatization, and feature extraction. These steps prepare the data for analysis and improve the model's efficiency and accuracy. The features like TF-IDF, Glove,Word2vec, BERT are extracted. The extracted features are fed into different classifiers.

Naive Bayes, SVM, Random Forest, Logistic Regression, and Decision Tree are traditional machine learning algorithms commonly used in fake news detection. These algorithms rely on statistical techniques



Raw dataset
- LIAR

NLP Techniques
- Text Preprocessing
- Feature Extraction
  - TF-TDF
  - Glove
  - word2Vec
  - BERT

Machine Learning Algorithms
- *Naïve Bayes*
- *Logistic Regression*
- *SVM*
- *Random Forest*
- *Decision Tree*

Deep Learning Algorithms
- CNN
- LSTM
- GRU

Evaluation
- Precision
- Recall
- F1-Score
- Accuracy

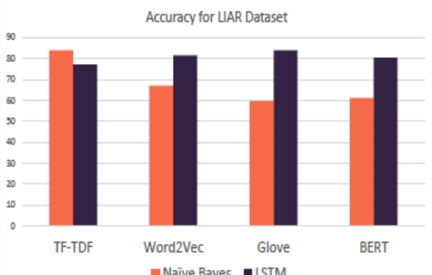
.

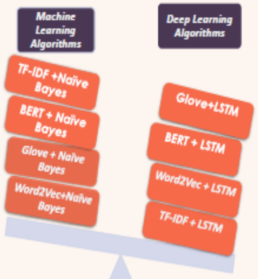
ML Algorithms

1. **Naive bayes** works on powerful nonaligned assumptions based on bayes theorem to classify the required text .
2. **Logistic regression** allows us to understand basic linear relationship between a dependent variable (i.e., fake or real news) and one or more independent variables
3. **SVM** is a binary classification algorithm that constructs a hyperplane to separate data points into different classes.
4. **Random forest** is an ensemble algorithm that constructs multiple decision trees and combines their predictions to improve accuracy.
5. **Decision tree** is a tree-based algorithm that creates a tree-like structure to model the decision-making process.

Deep Learning Algorithms

1. **CNNs:** Convolutional Neural Networks are commonly used in image recognition but can also be applied to text data by extracting features through convolutional layers.
2. **Long Short-Term Memory (LSTM):** This has the capability to capture the temporal dependencies in the text. It works well with long sequences of text and can handle variable input length and it also belongs to RNN family.
3. **Gated Recurrent Unit (GRU):** This algorithm is a variant of LSTM that is computationally more efficient. It has fewer parameters than LSTM but can still capture the temporal dependencies in the text.





Using TF-IDF, Naïve Bayes achieves a higher accuracy of 83.92%, while the LSTM model achieves an accuracy of 77.05%. Using Word2Vec embeddings, the LSTM model achieves a higher accuracy of 81.43% compared to Naïve Bayes with an accuracy of 67.02%. Using Glove embeddings, the LSTM model achieves a much higher accuracy of 83.99%, while Naïve Bayes achieves an accuracy of 59.71%. Using BERT embeddings, the LSTM model achieves a slightly lower accuracy of 80.37%, while Naïve Bayes achieves an accuracy of 61.21%. Overall, we can observe that the LSTM model tends to perform better than Naïve Bayes for all the feature extraction techniques except for the TF-IDF technique. [1]

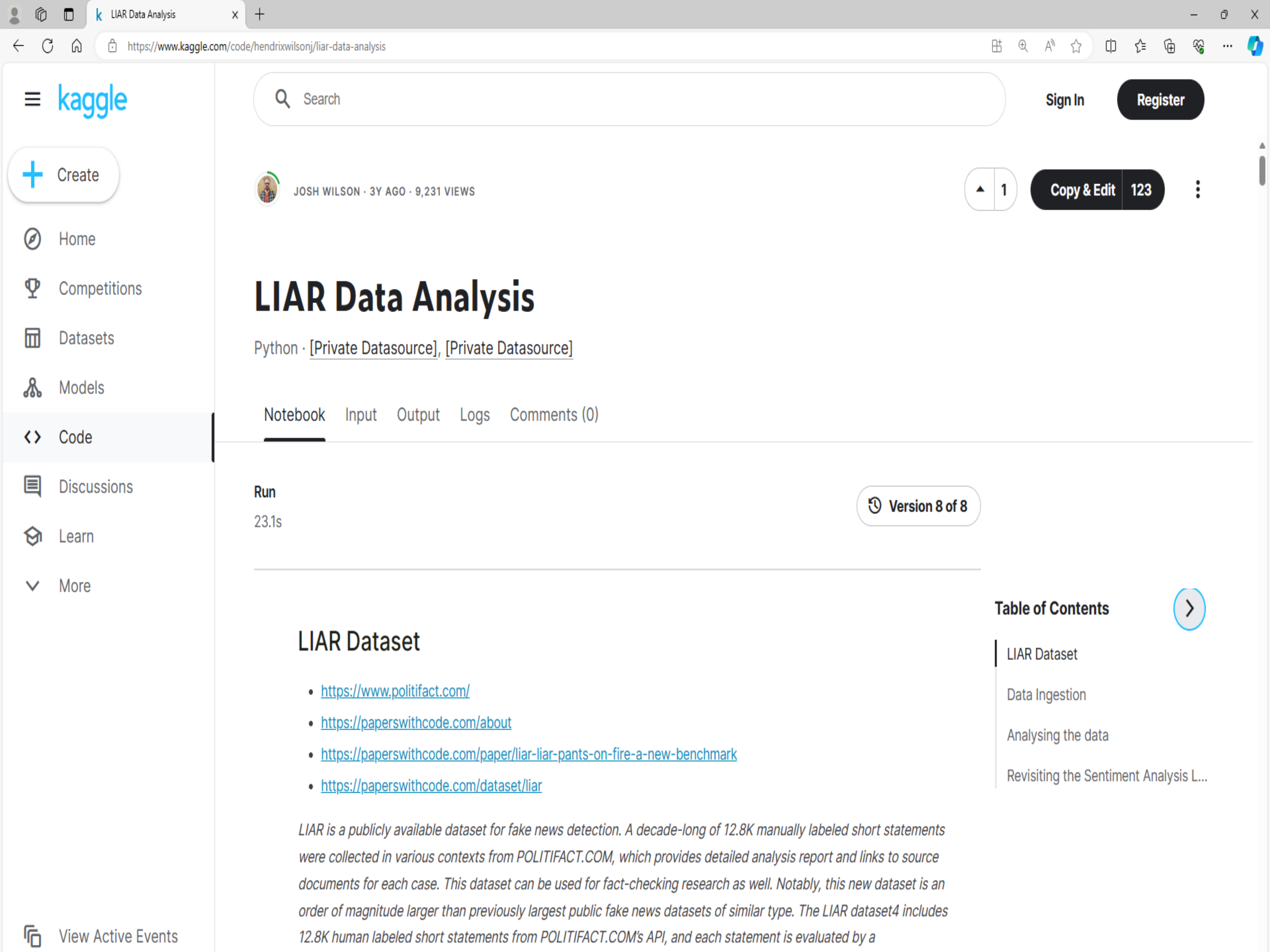
Conclusion

For LIAR dataset, when it comes to detecting fake news using natural language processing techniques, we can observe that the deep learning models with dense word embeddings tend to perform better than the traditional machine learning model. This is especially true when using word2vec, glove embeddings, and BERT, which allow these algorithms to capture complex relationships and patterns within the data. However, when using TF-IDF embedding, traditional machine learning algorithms like Naive Bayes tend to perform better. This suggests that the choice of embedding and an appropriate algorithm together plays a crucial role in determining the accuracy of the algorithm. [1]

For other datasets, using BERT embedding technique along with deep learning algorithms has shown better accuracy compared to traditional embedding models like TF-IDF, word2vec, Glove. [2]

References

1. https://www.mdpi.com/2227-7390/11/3/508
2. https://link.springer.com/article/10.1007/s11042-020-10183-2?ArticleAuthorOnlineFirst_20210110&error=cookies_not_supported&code=6ce56298-7be6-49dc-90d7-5764a17db66b
3. https://www.kaggle.com/datasets/csmalarkodi/l



LIAR Data Analysis

Python · [Private Datasource], [Private Datasource]

Run

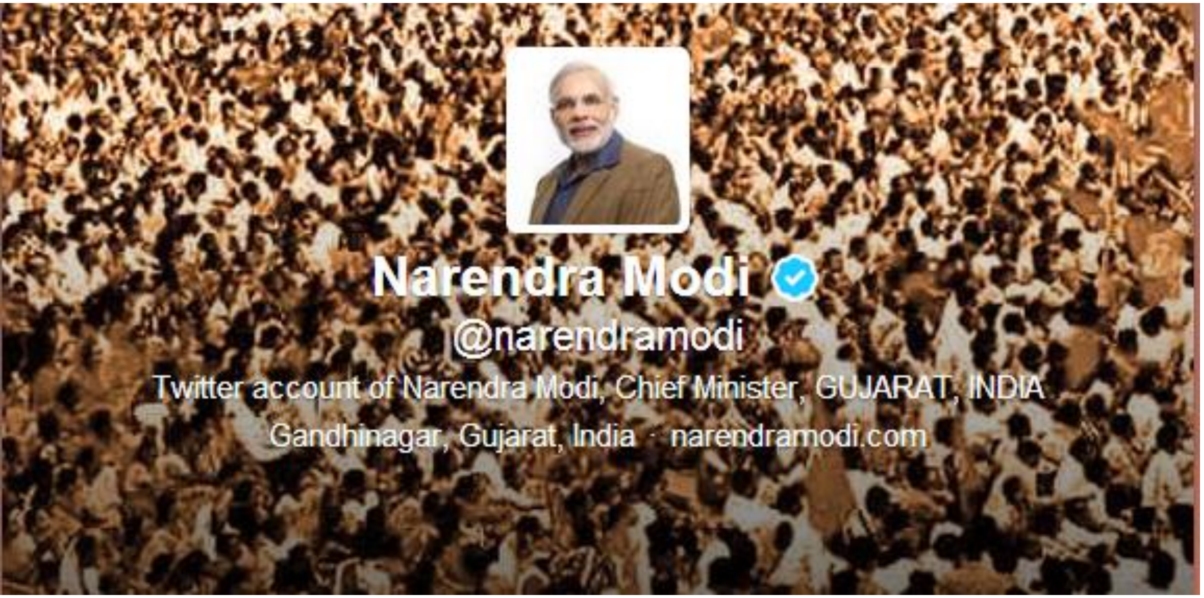
23.1s

LIAR Dataset

- https://www.politifact.com/
- https://paperswithcode.com/about
- https://paperswithcode.com/paper/liar-liar-pants-on-fire-a-new-benchmark
- https://paperswithcode.com/dataset/liar

LIAR is a publicly available dataset for fake news detection. A decade-long of 12.8K manually labeled short statements were collected in various contexts from POLITIFACT.COM, which provides detailed analysis report and links to source documents for each case. This dataset can be used for fact-checking research as well. Notably, this new dataset is an order of magnitude larger than previously largest public fake news datasets of similar type. The LIAR dataset4 includes 12.8K human labeled short statements from POLITIFACT.COM's API, and each statement is evaluated by a

Who will be the next Prime Minister of India?



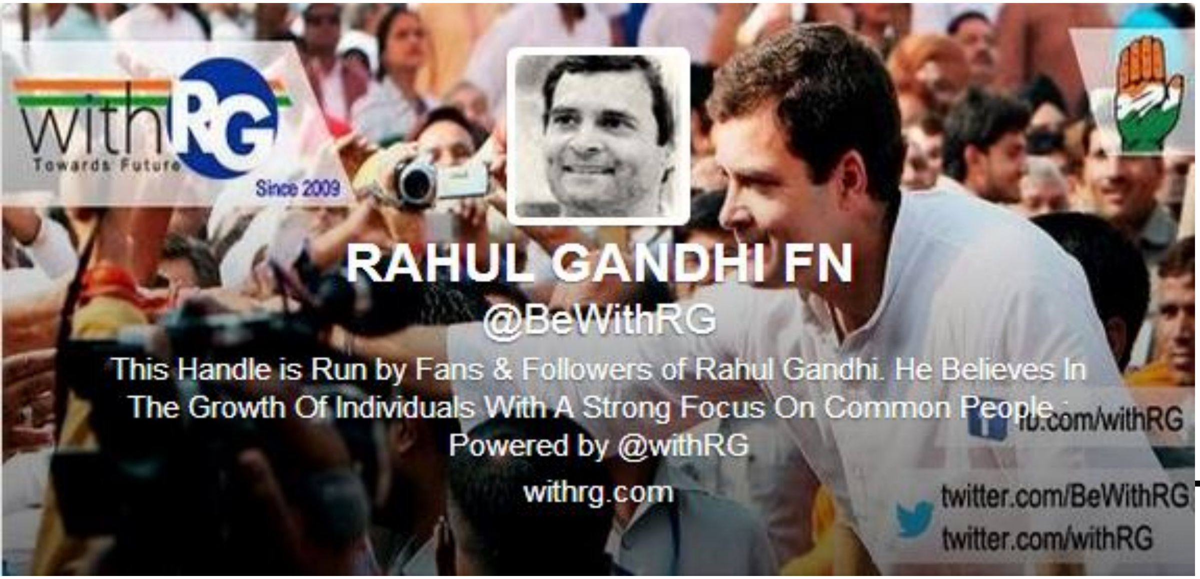
Narendra Modi

@narendramodi

Twitter account of Narendra Modi, Chief Minister, GUJARAT, INDIA

Gandhinagar, Gujarat, India · narendramodi.com

Use Sentiment Analysis techniques on twitter to gauge sentiments for next General Elections of India



RAHUL GANDHI FN

@BeWithRG

This Handle is Run by Fans & Followers of Rahul Gandhi. He Believes In The Growth Of Individuals With A Strong Focus On Common People ; Powered by @withRG

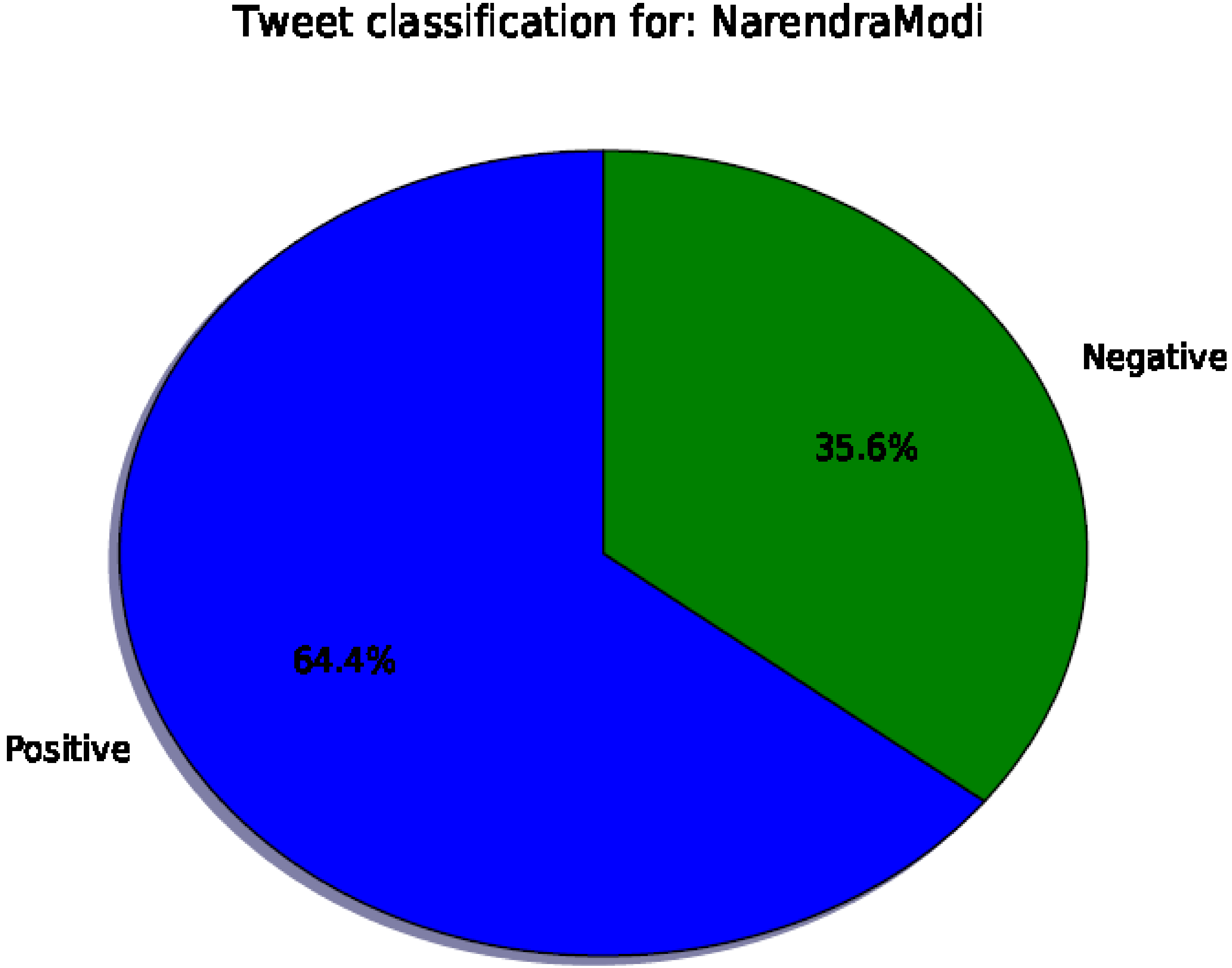
withrg.com

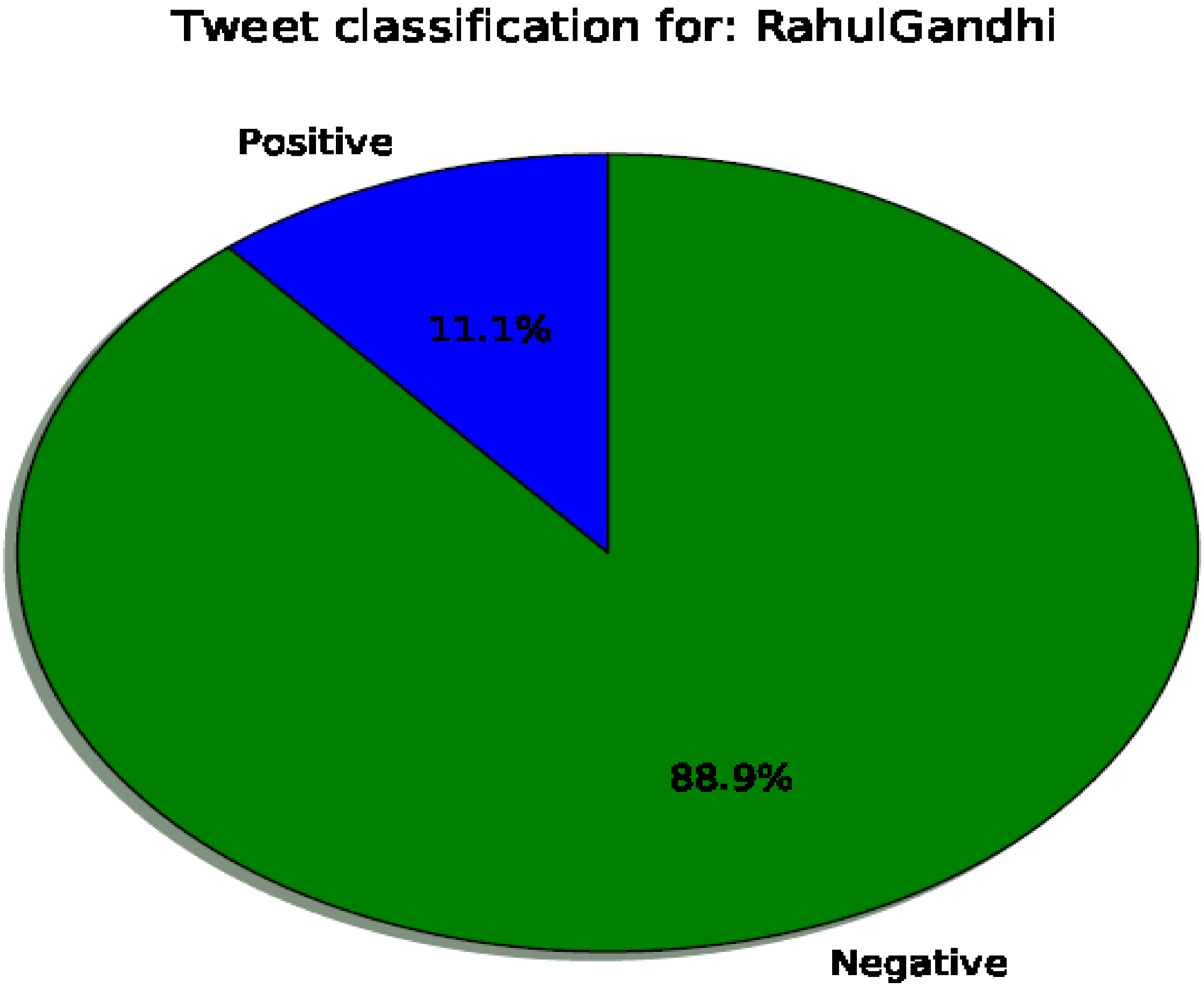
fb.com/withRG

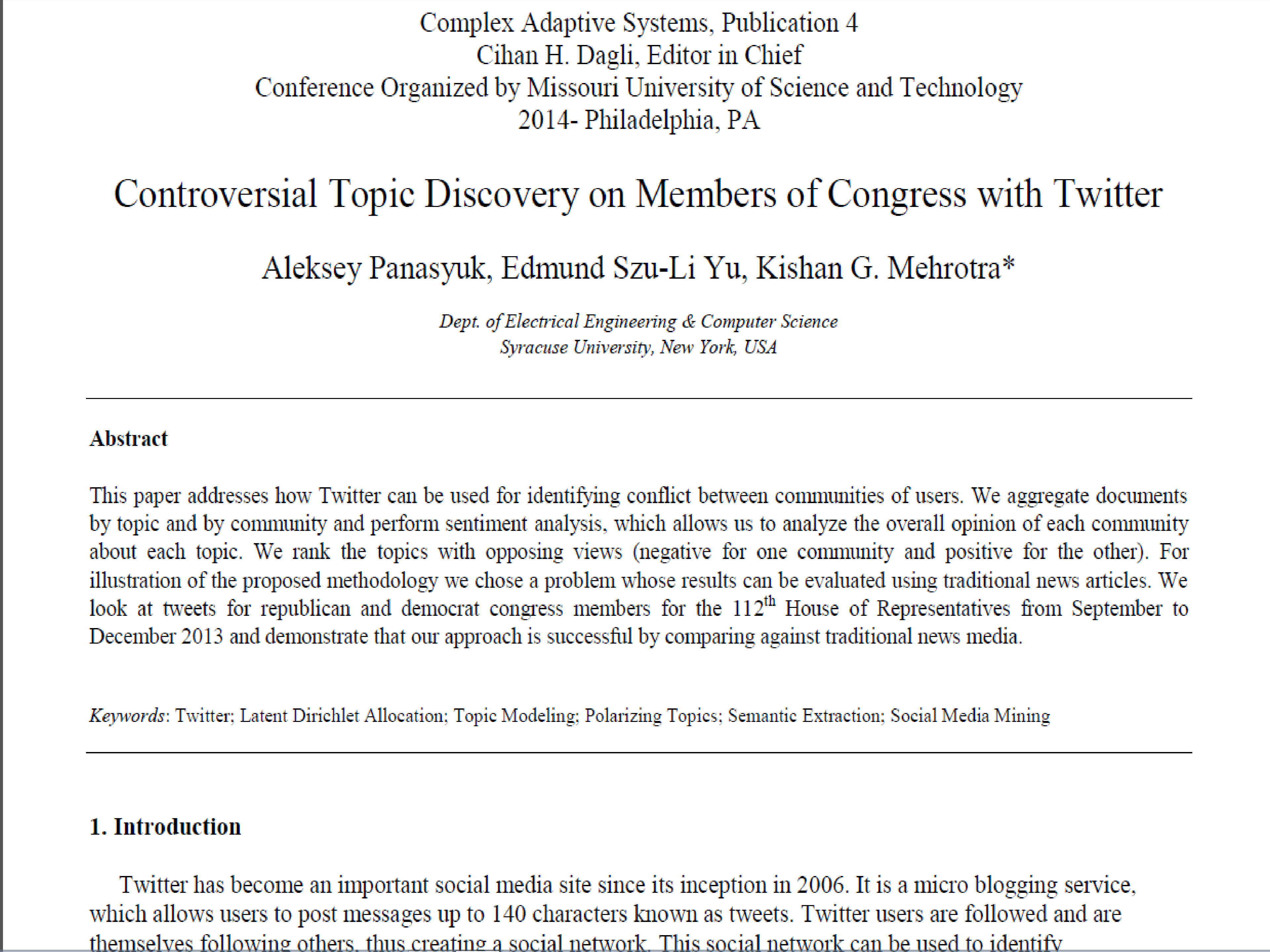
twitter.com/BeWithRG

twitter.com/withRG

Prasoon Dilip Pandya


Tweet classification for: NarendraModi
Negative
35.6%
64.4%
Positive


Tweet classification for: RahulGandhi
Positive
11.1%
88.9%
Negative



Complex Adaptive Systems, Publication 4
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2014- Philadelphia, PA

Controversial Topic Discovery on Members of Congress with Twitter

Aleksey Panasyuk, Edmund Szu-Li Yu, Kishan G. Mehrotra*

Dept. of Electrical Engineering & Computer Science
Syracuse University, New York, USA

Abstract

This paper addresses how Twitter can be used for identifying conflict between communities of users. We aggregate documents by topic and by community and perform sentiment analysis, which allows us to analyze the overall opinion of each community about each topic. We rank the topics with opposing views (negative for one community and positive for the other). For illustration of the proposed methodology we chose a problem whose results can be evaluated using traditional news articles. We look at tweets for republican and democrat congress members for the 112th House of Representatives from September to December 2013 and demonstrate that our approach is successful by comparing against traditional news media.

Keywords: Twitter; Latent Dirichlet Allocation; Topic Modeling; Polarizing Topics; Semantic Extraction; Social Media Mining

1. Introduction

Twitter has become an important social media site since its inception in 2006. It is a micro blogging service, which allows users to post messages up to 140 characters known as tweets. Twitter users are followed and are themselves following others, thus creating a social network. This social network can be used to identify

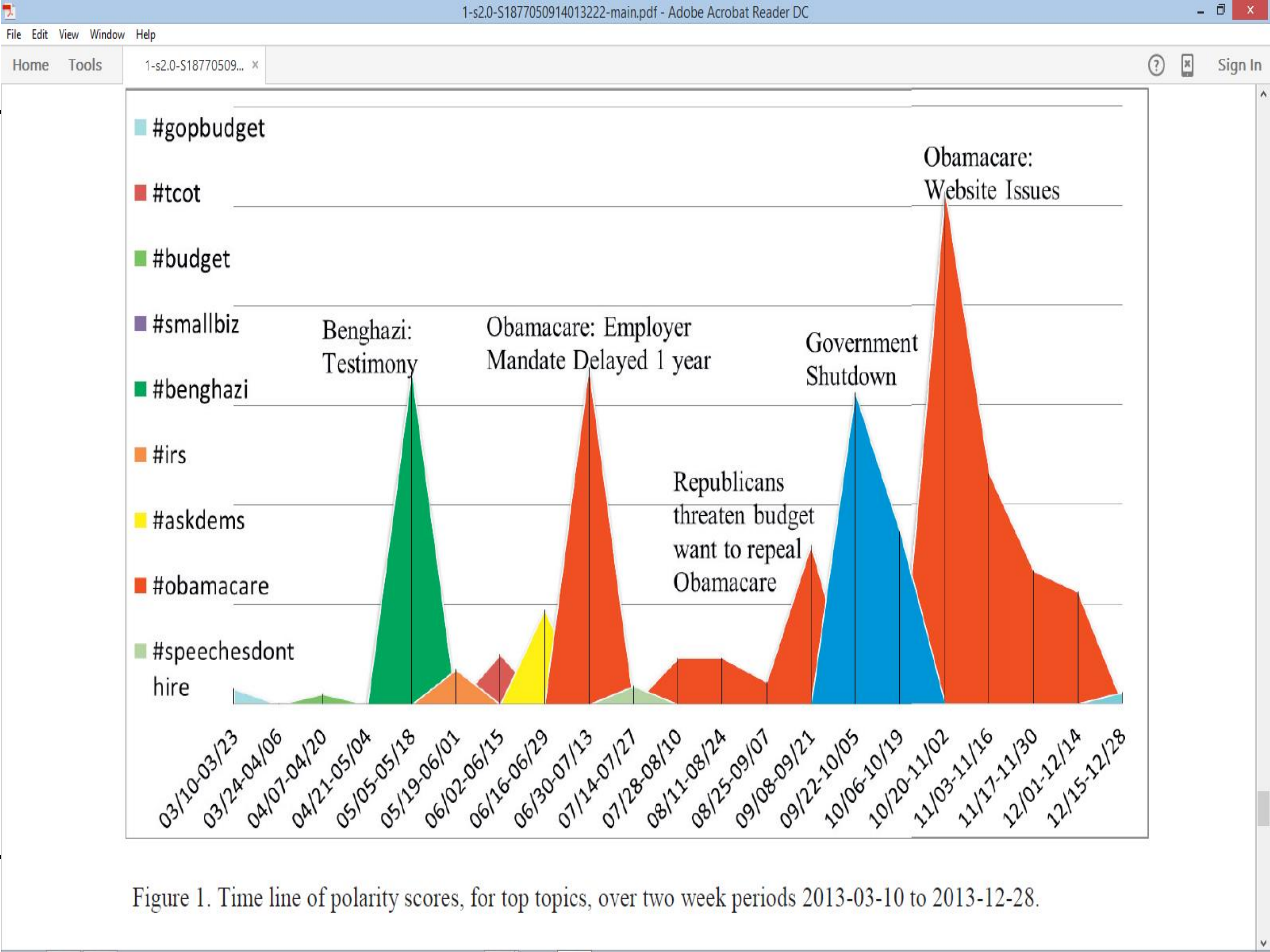
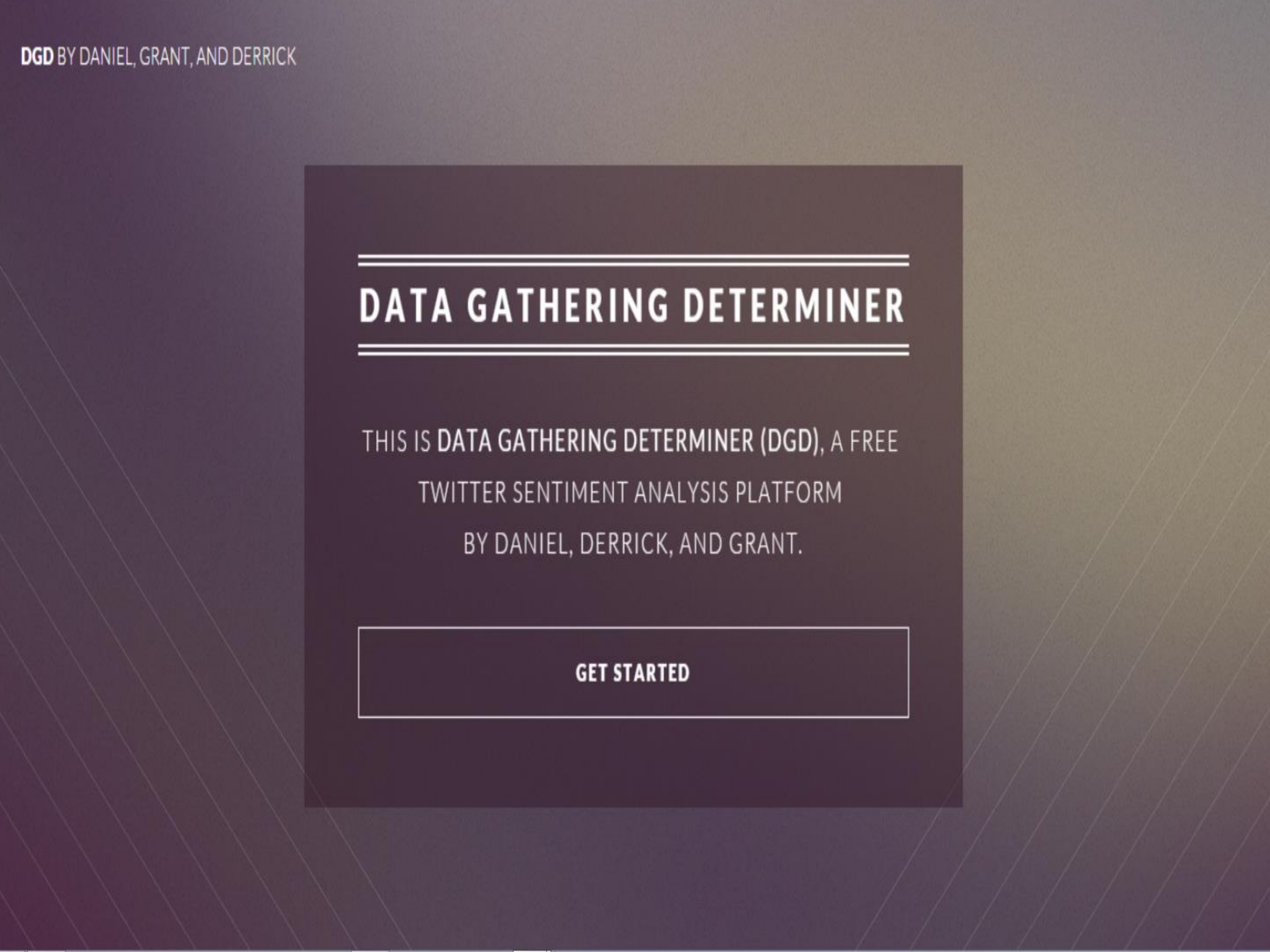


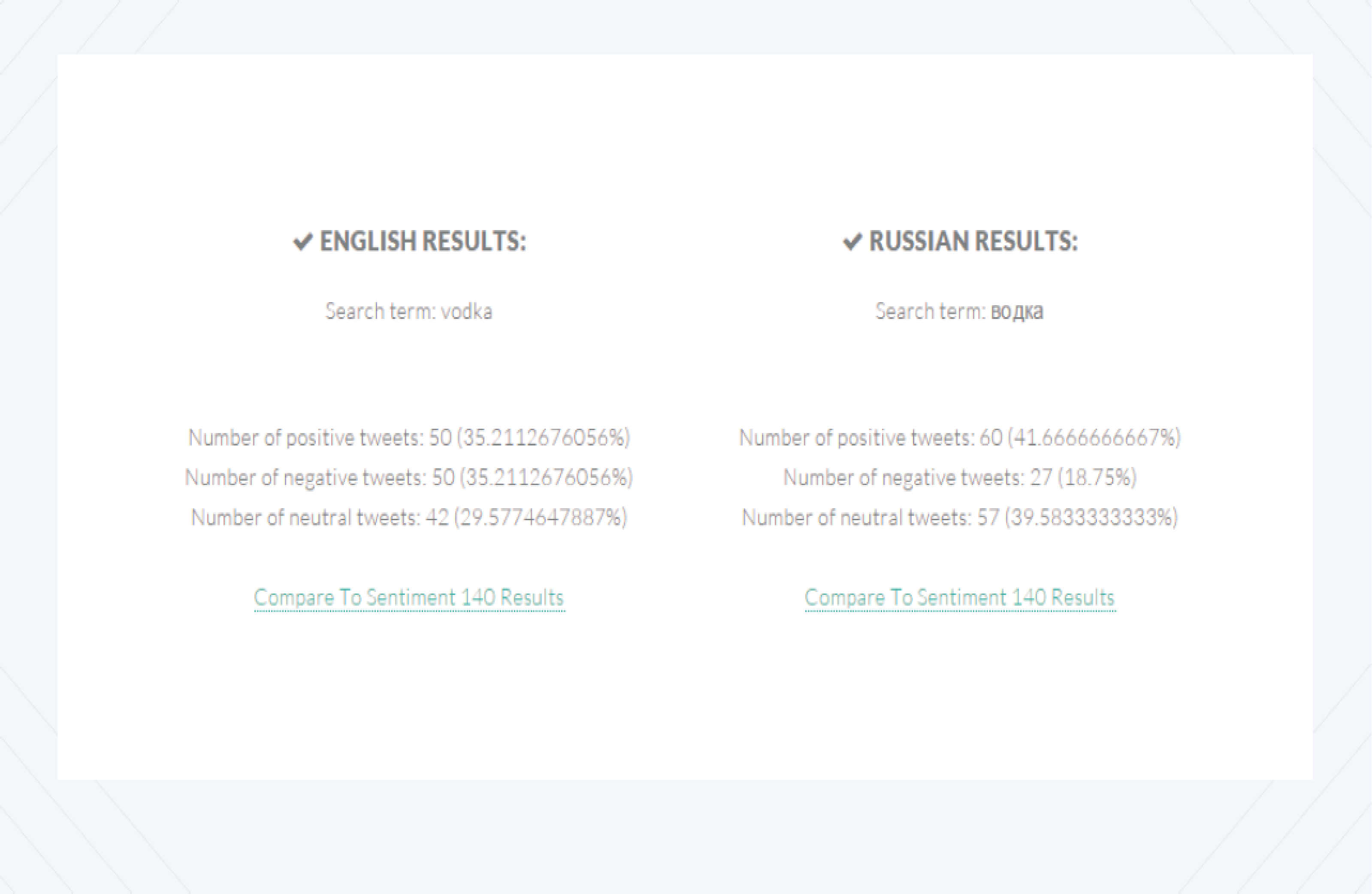
Figure 1. Time line of polarity scores, for top topics, over two week periods 2013-03-10 to 2013-12-28.



DATA GATHERING DETERMINER

THIS IS **DATA GATHERING DETERMINER (DGD)**, A FREE TWITTER SENTIMENT ANALYSIS PLATFORM BY DANIEL, DERRICK, AND GRANT.

Results page



✓ ENGLISH RESULTS:

Search term: vodka

Number of positive tweets: 50 (35.2112676056%)

Number of negative tweets: 50 (35.2112676056%)

Number of neutral tweets: 42 (29.5774647887%)

Compare To Sentiment 140 Results

✓ RUSSIAN RESULTS:

Search term: водка

Number of positive tweets: 60 (41.6666666667%)

Number of negative tweets: 27 (18.75%)

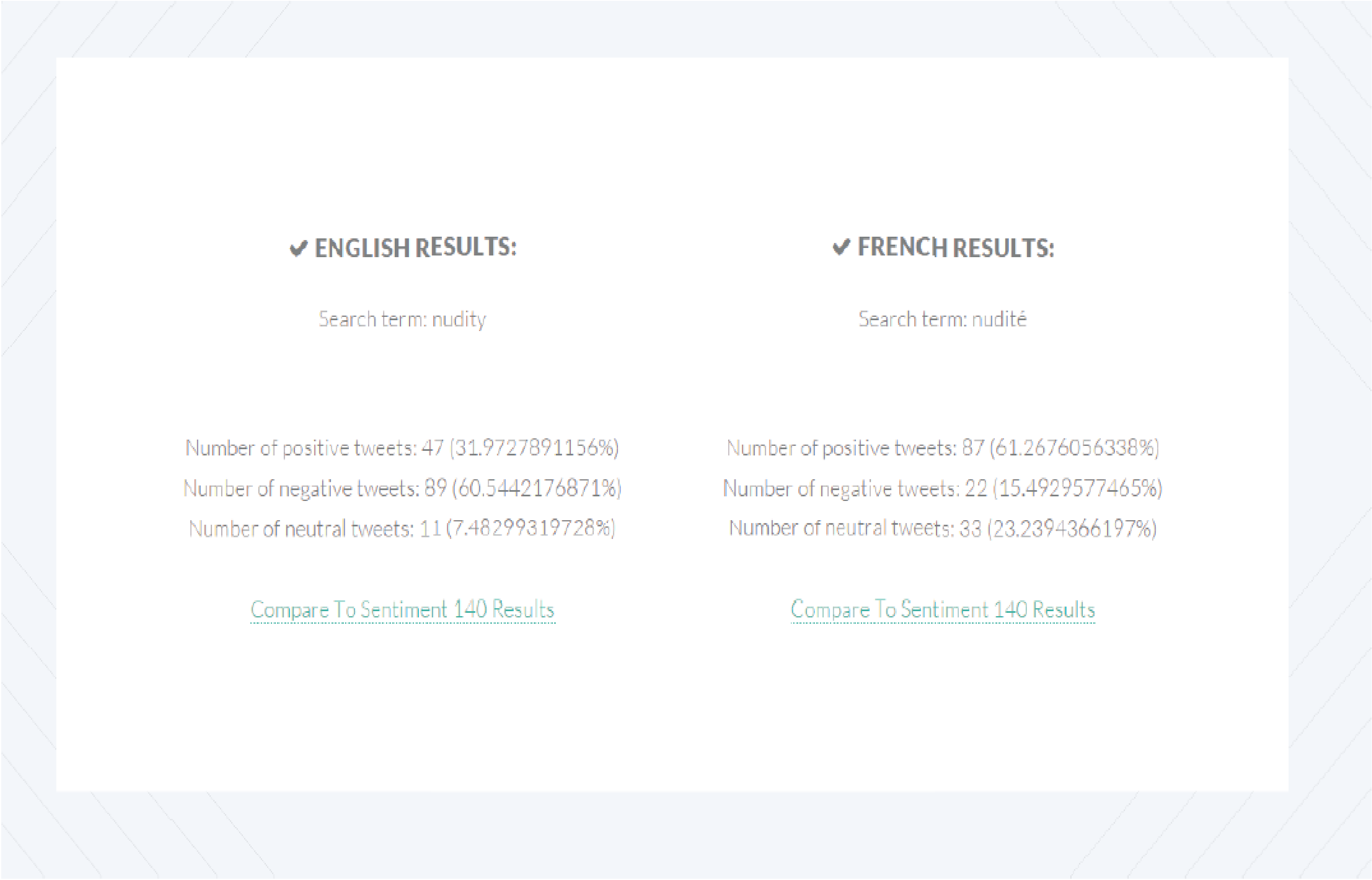
Number of neutral tweets: 57 (39.5833333333%)

Compare To Sentiment 140 Results

Prediction: Russian people generally are more tolerant of vodka, so they should have more positive tweets than negative.

Result: Mostly met expectations, though not too extremely.

Results page



✓ ENGLISH RESULTS:

Search term: nudity

Number of positive tweets: 47 (31.9727891156%)
Number of negative tweets: 89 (60.5442176871%)
Number of neutral tweets: 11 (7.48299319728%)

Compare To Sentiment 140 Results

✓ FRENCH RESULTS:

Search term: nudité

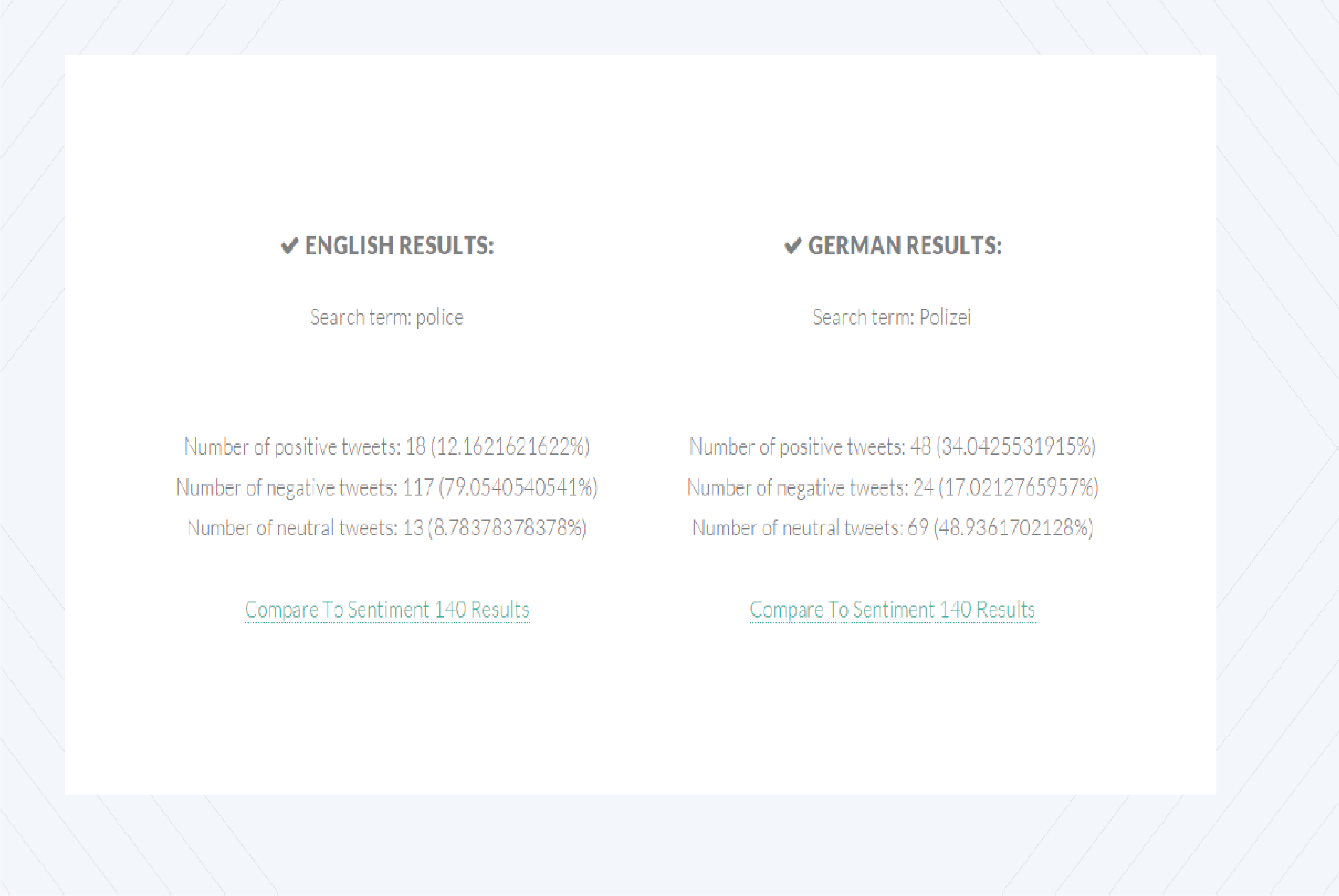
Number of positive tweets: 87 (61.2676056338%)
Number of negative tweets: 22 (15.4929577465%)
Number of neutral tweets: 33 (23.2394366197%)

Compare To Sentiment 140 Results

Prediction: Nudity is seen negatively in english speaking countries, whereas french speakers are likely to be more okay with nudity due to cultural reasons.

Result: Matches our prediction

Results page



✔ ENGLISH RESULTS:

Search term: police

Number of positive tweets: 18 (12.1621621622%)
Number of negative tweets: 117 (79.0540540541%)
Number of neutral tweets: 13 (8.78378378378%)

Compare To Sentiment 140 Results

✔ GERMAN RESULTS:

Search term: Polizei

Number of positive tweets: 48 (34.0425531915%)
Number of negative tweets: 24 (17.0212765957%)
Number of neutral tweets: 69 (48.9361702128%)

Compare To Sentiment 140 Results

Prediction: Police has had a very poor image in the english media lately due to recent incidents. There will be a lot of negative tweets.

Result: Matches our prediction

Semantic analysis of self-isolation tweets in the USA and Russia

DARIA SINITSYNA

COMPUTATIONAL LINGUISTICS

Introduction

- While the USA took *a long time to adequately respond to the pandemic*, the government now seemingly does everything in their power to maintain the economy and health of the society
- In Russia the government does not officially name the situation in the country "*quarantine*", instead only saying "*self-isolation*" so they would not have to declare a state of emergency
- Knowing people's opinions about self-isolation would be extremely valuable to understand whether the different states in the USA and the government of Russia are doing *a proficient job to maintain heath and well-being of the citizens*
- To acquire this information, we would need to perform **sentiment analysis on social media data**

Training the models

ENGLISH LANGUAGE MODEL

Learning data:

1. Sentiment140

2. SemEval-2017 Task 4

3. Collected data from reputable news sources to account for the neutral label

RUSSIAN LANGUAGE MODEL

Learning data:

1. Positive and negative data from https://study.mokoron.com/

2. Collected data from reputable news sources to account for the missing neutral label

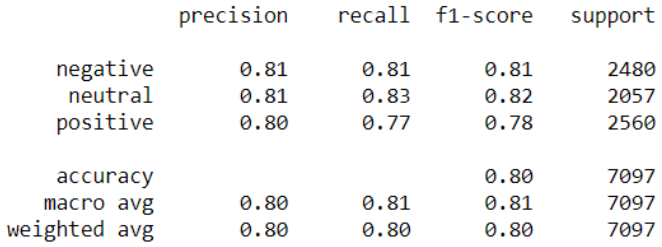
Results of training

ENGLISH LANGUAGE MODEL

TfIdf vectorization:

1. without omitting the stopwords
2. using both unigrams and bigrams
3. using min_df = 2

SVM model performed best:



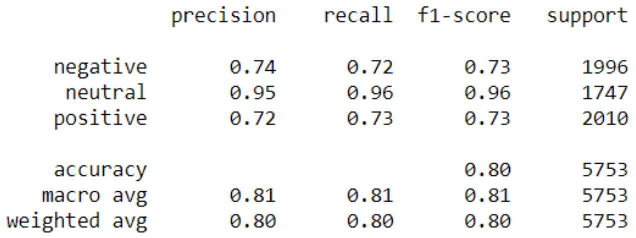
```
              precision    recall  f1-score   support

    negative       0.81      0.81      0.81      2480
     neutral       0.81      0.83      0.82      2057
    positive       0.80      0.77      0.78      2560

    accuracy                           0.80      7097
   macro avg       0.80      0.81      0.81      7097
weighted avg       0.80      0.80      0.80      7097
```

RUSSIAN LANGUAGE MODEL

TfIdf vectorization:

1. without omitting the stopwords
2. using both unigrams and bigrams
3. using min_df = 2

SVM model performed best (but **overfitted** on neutral):



```
              precision    recall  f1-score   support

    negative       0.74      0.72      0.73      1996
     neutral       0.95      0.96      0.96      1747
    positive       0.72      0.73      0.73      2010

    accuracy                           0.80      5753
   macro avg       0.81      0.81      0.81      5753
weighted avg       0.80      0.80      0.80      5753
```

Collecting self-isolation data

- For Russia, we use the “самоизоляция” (self-isolation) keyword and are only looking for the tweets in Russian - 5520 tweets

- For the USA, we look at the following states (hardest hit by COVID-19) – 3772 tweets:

1. New York
2. New Jersey
3. Massachusetts
4. Illinois
5. California
6. Pennsylvania
7. Michigan
8. Florida
9. Texas
10. Louisiana
11. Georgia
12. Washington

Opinion mining for American tweets

California	neutral 92 negative 60 positive 47	Michigan	neutral 240 positive 94 negative 65	Massachusetts	neutral 139 positive 39 negative 34
Florida	neutral 117 positive 44 negative 30	New York / New Jersey / Pennsylvania	neutral 605 positive 167 negative 155	NYC	neutral 414 positive 107 negative 84
Georgia	neutral 17 negative 17 positive 11	Texas	neutral 189 negative 70 positive 64	Washington	neutral 330 positive 121 negative 72
Louisiana	neutral 78 positive 31 negative 24	Illinois	neutral 132 positive 47 negative 41		

Most informative words in American tweets

Negative	Neutral	Positive
Acquired immunity	Mental health	Balcony
Get back to work	Infected	Music
Order	Confirmed	Cat
Work	Positive (test)	Help
Go (out)	Essential (workers)	Forward
Sorry	Food	Perfect
	Groceries	Friends
	Cleanliness	Learned
	Routine	
	Sane	
	Mandatory	

Opinion mining for Russian tweets

1. As for the positive label, there are several more informative words: *friend, good (day/friend), mother/mom, store*. Mostly, people are trying to stay positive (or sarcastically pretend to stay positive) and are talking about their current situation in life with this crisis and how they keep sane.

2. When looking at the neutral label, noticeably, there are such words as: *virus, Moscow, government, national, Putin, country*, etc. As with the USA, mostly people neutrally discuss the news about the prolonged self-isolation order, about governmental response, about Moscow orders for quarantine action.

3. In the negative category, there are such words as: obscene lexicon, *get out, why, hard, money, bad, last money, critically,* etc. Here, the negative result is prompted by the lack of adequate response to the pandemic by the government and no economic relief packages that exist in the USA. People are out of jobs, cannot apply for unemployment, are losing their businesses and “*are on their last money*”.

Conclusion

1. Mostly, in both countries the results for the positive label looked similar*: people were tweeting about their ways of keeping their heads up, staying positive and in touch with their loved ones, daily routines that had to change, etc*.

2. Looking at the neutral label, we can see that in both countries the tweets were about *general information on self-isolation*; **however**, the USA data seems to be more informative because of the Russian data bias towards news topics, as Russian neutral-labelled data consisted mostly of words related to news and government

3. When it came to the negative label, the results were, again, similar, as they brought up people's *dissatisfaction with the governmental response towards the pandemic*