ENTHUSIASM AND SUPPORT:
ALTERNATIVE SENTIMENT CLASSIFICATION FOR SOCIAL MOVEMENTS ON SOCIAL MEDIA

Shubhanshu Mishra, Sneha Agarwal, Jinlong Guo, Kirstin Phelps, Johna Picco, Jana Diesner
{ smishra8, sagarwa8, jguo24, kphelps, picco2, jdiesner }@illinois.edu

Introduction

Sentiment analysis has for a long time been looked as a robust measure of extracting emotion from language constructs. Network analysis can be combined with sentiment classification to identify entities (users, hashtags) from twitter corpus which represent a specific class of support and enthusiasm towards a social cause. Our resulting computational solution can help organizations involved with social causes to disseminate messages in a more informed and effective fashion; potentially leading to greater impact.

Workflow

We set out to solve the defined problem by using the following workflow which will result in finally devising a technique to classify users based on sentiment in a social media network.

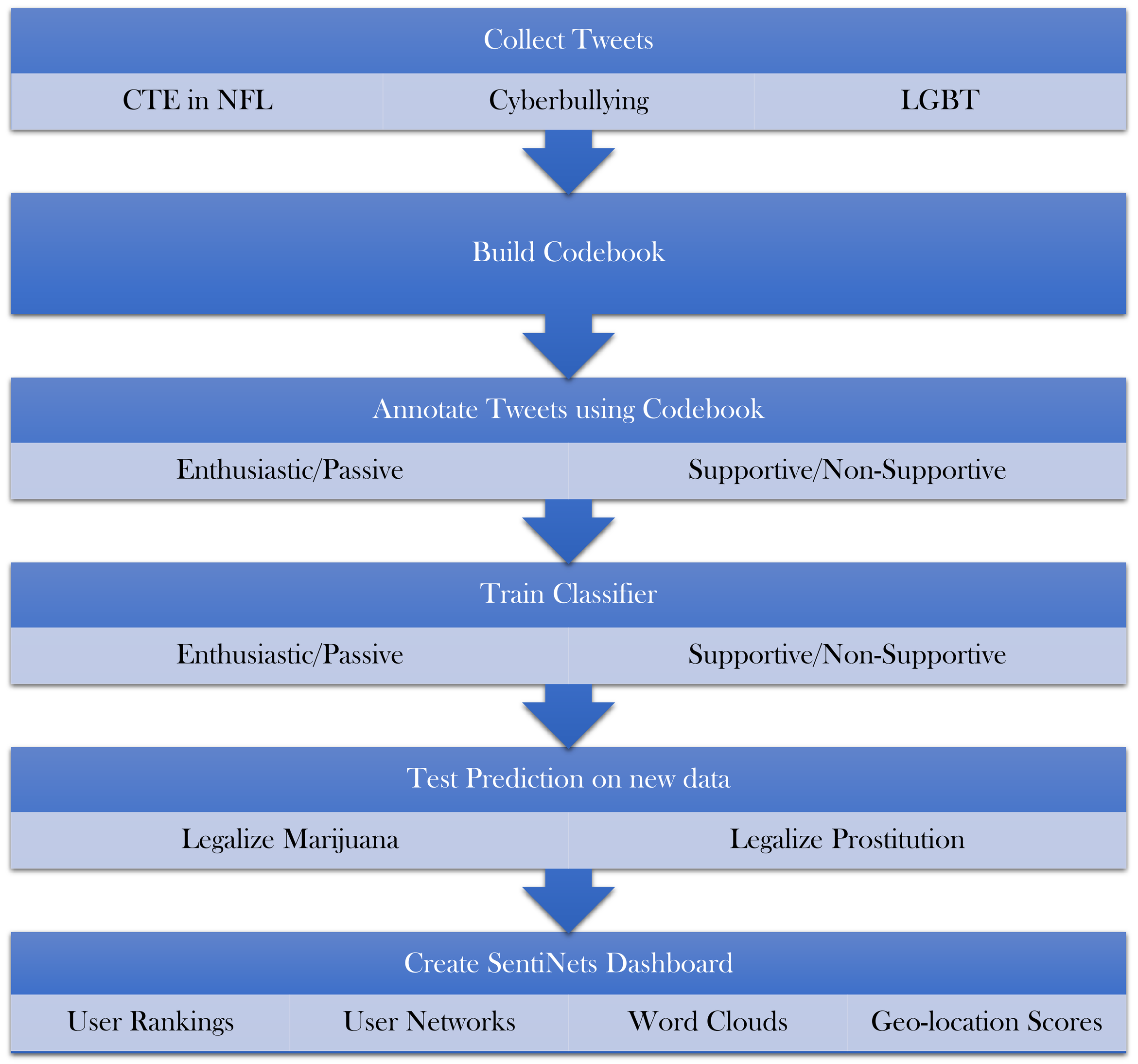


Fig 1. SentiNets Workflow

Tweet Corpus

1500 Tweets collected for the following social causes. The corpus didn't have duplicate tweets and had only tweets with length greater than 3 words.

- **Lesbian Gay Bisexual Transgender [LGBT]**
- **Concussions in National Football League [CTE in NFL]**
- **Cyberbullying**

Methods

Classification Schema

Understanding the needs for identifying users/tweets for social causes the classification of 2 orthogonal classes was found to be best suited. This classification schema allows us to move beyond the positive and negative sentiment classification of tweets to a more audience identification centric approach.

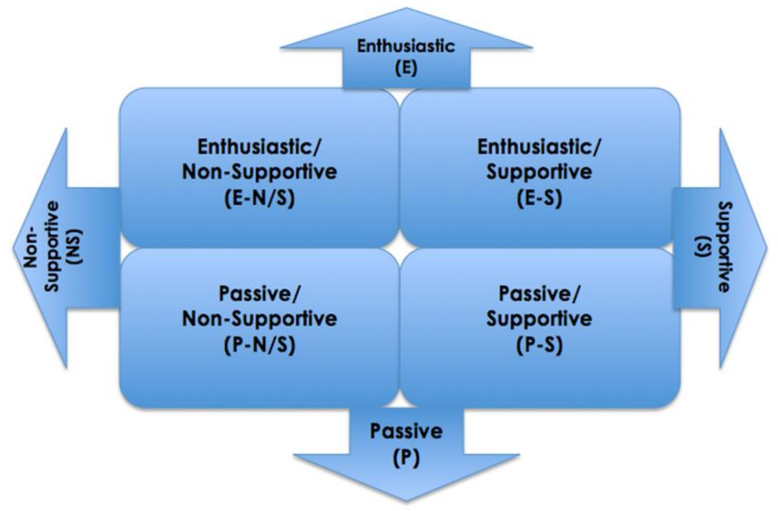


Fig 2. Sentiment classification schema

Codebook Generation

We created a codebook which will help us in generating a training corpus for our classifier for each of the following classes. The codebook was used to hand code the 1500 tweet corpus along the 2 orthogonal classification schemas. We avoided using context based knowledge for getting the coding done so as to remove personal opinions from the coding scheme.

For Non-Supportive class we considered the case where the tweets were either directly against the cause or just spreading negative information about the cause. We merged these two cases to build the Non-supportive class as the corpus had very few tweets which were directly against the cause.

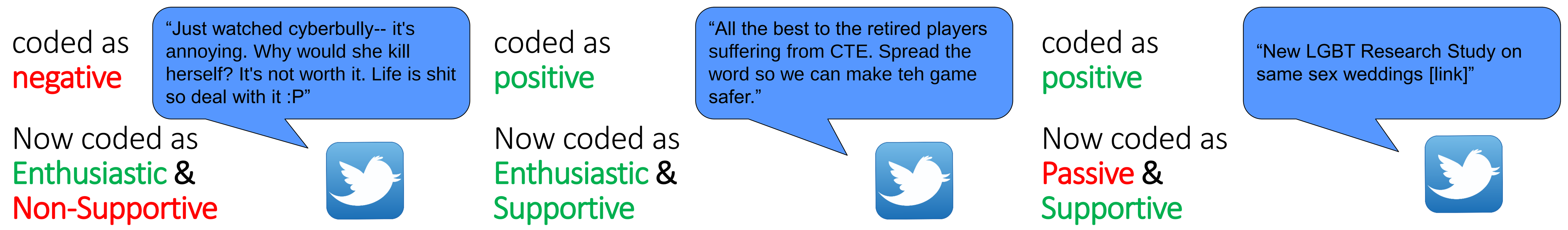
Classifier Training

Once we had the training corpus we decided to train a ***Linear Support Vector Machine (SVM)*** based classifier. The classifier was trained using the features shown in the table below. We used 10 fold cross validation to train the classifier and report the accuracies.

# of Emoticons	*# of URLS*	*# of Mentions*	*# of Hashtags*
Word Features	*# of Double Quotes*	*Length of Tweets*	

Results

Category	Inter Coder Reliability	Accuracy (SVM)
Enthusiastic v/s Passive	**93 %**	**79 %**
Supportive v/s Non - Supportive	**85 %**	**77 %**



Codebook Results

While building the codebook we observed the following key issues related to classification based on our scales:

- It was found that less than 10% of the people speak openly against a cause in a public platform like social media.
- Supportive/Non-Supportive scale was found to be harder to code consistently as it does require some subjective knowledge as compared to enthusiastic/passive scale

Classifier Testing

We tested our classifier on two new topics (viz. ***"Legalize Marijuana" and "Legalize Prostitution"***) and got very good results for the Enthusiastic v/s Passive Scale. The Supportive v/s Non-Supportive case was influenced by the nuances in sentiment classification and needs more improvement.

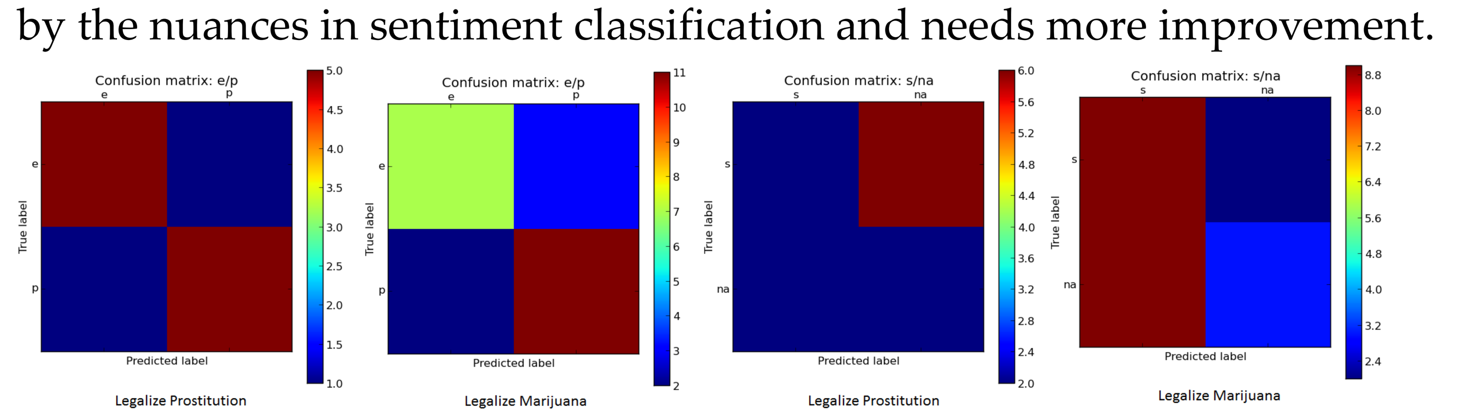
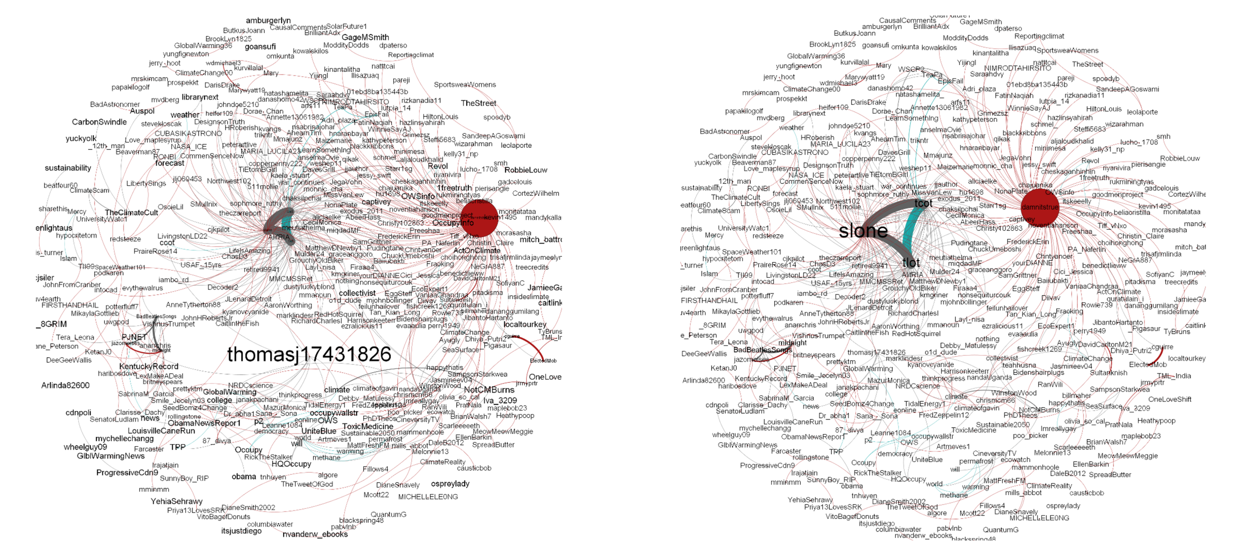


*Fig 3. Confusion Matrices for sentiment classes for **Legalize Marijuana** and **Legalize Prostitution***

Sentiment based Networks

We use out sentiment classifiers to classify each tweet and assign the sentiment of the tweet to each user and hashtag involved with the tweet. We aggregate the total sentiment scores for each class for each user and hashtag to generate the node statistics for our network. Each edge in our network is entities occurring together.



Node Color: *HashTags or User* — ***Node Size:*** *Occurrence* — ***Label Size:*** *Enthusiasm Measure*

Node Color: *HashTags or User* — ***Node Size:*** *Occurrence* — ***Label Size:*** *Support Measure*

Fig 4. Sentiment based Networks for Global Warming

Label	Type	Weight	Tweet_Count	E_Count	P_Count	S_Count	NS_Count	EP_Class	SNS_Class	Degree	Top in Class
Sorted by Weights											
damnitstrue	USER	91	0	0	91	0	91	PASSIVE	NON_SUPPORTIVE	92	
slone	USER	51	0	0	51	50	1	PASSIVE	NON_SUPPORTIVE	54	SUPPORTIVE
tcot	HASHTAG	50	0	1	49	47	3	PASSIVE	NON_SUPPORTIVE	56	SUPPORTIVE
Sorted by Number of Tweets											
ElectedMob	USER	9	9	0	9	5	4	PASSIVE	NON_SUPPORTIVE	2	
thomasj17431826	USER	7	6	6	1	1	6	ENTHUSIASTIC	SUPPORTIVE	11	ENTHUSIASTIC
NotCMBurns	USER	6	6	4	2	6	0	ENTHUSIASTIC	SUPPORTIVE	4	ENTHUSIASTIC

In above generated network for the cause **"global warming"** using our tool **SentiNets** we see that there are 2 individuals who are prominent in each of the classifications. We also see that majority of the tweets tend to be passive and non-supportive. A reason for this is that most people tend to share links or news items on social causes. From the above table we infer that the most enthusiastic user has sent out a lot of tweets whereas the most supportive user has not but was retweeted the most and occurs with the top most tweeted hashtags. The most retweeted user is the one with the most passive and non-supportive tweet.

Acknowledgements

Our tool SentiNets is made available in the text network analysis tool called ConText http://context.lis.illinois.edu
Our classifiers were trained using Weka and the visualization was done in Gephi.
More details about SentiNets can be seen at the official page: http://people.lis.illinois.edu/~smishra8/sentinets.php

ILLINOIS
GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN