ILLINOIS INSTITUTE OF TECHNOLOGY

INFLUENCE OF TIE STRENGTH ON HOSTILITY IN SOCIAL MEDIA

Bahar Radfar
Master's Thesis Defense
Spring 2019

Thesis Committee:
Dr. Aron Culotta
Dr. Mustafa Bilgic
Dr. Xiaoqian Li

Agenda

- Problem Statement
- Data Collection and Annotation
- Data Analysis and Feature Extraction
- Classification Model
- Evaluation and Results
- Future work

Toxicity on Social Media

- Over the past decade, the grown portion of human communication is occurring over social media. (From 0.97 B to 2.77 B users [1])
- Unfortunately, the amplification of social connectivity also includes the amplification of the negative aspect of society, leading to a huge amount of hostility.

[1] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

What does the **hostility** mean?

- Hostility comes from the Latin origin of "hostilis", which mean unfriendliness, opposition, and acts of warfare.
- A hostile conversation:
  - Includes one or more words that represent harassing, bullying, threatening, and sexual language towards an individual or a group.

Forbidden words

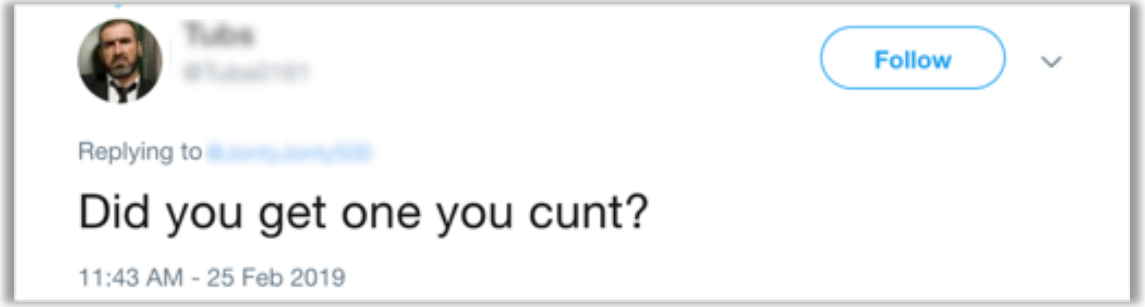
- Differences between cyberbullying and traditional bullying:
  - Cyberbullying can happen anywhere, anytime
  - Cyberbullying incidents can go viral
  - **Anonymity: the internet protects the bully**

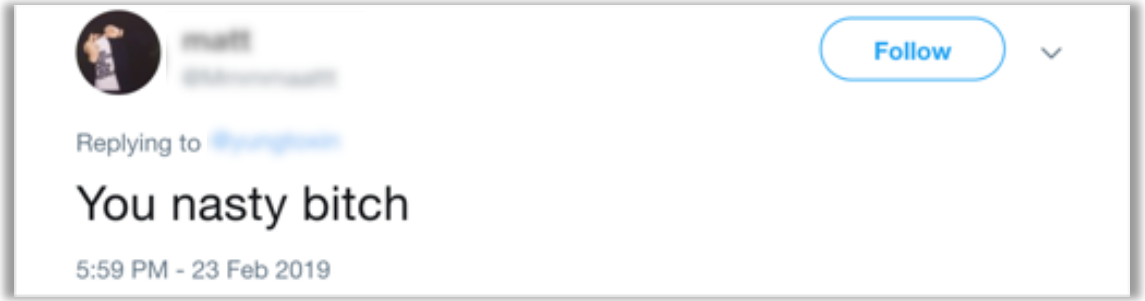
- **Why do not** we define a set of forbidden words?

Problem statement

What do you think about these tweets' level of hostility?

Can the Friendship change the negative power of a hostile word?





YES, it can!


Feb 25
14500 applications for an away game at Wolves. Absolutely no other club like it.
2
5
34
Follow
Replying to
Did you get one you cunt?
11:43 AM - 25 Feb 2019
1
Feb 25
Replying to
No.
1
Feb 25
Did you?
1
Feb 25
Aye

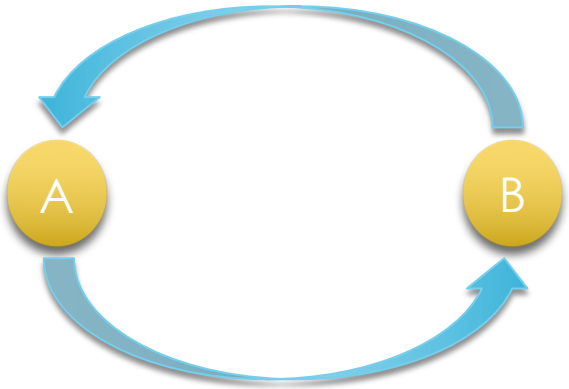
Proposed Solution

- Define relationship groups
- Categorize the data based on the relationships
- Analyze each group's data in term of hostility

Relationship Categories

No Friendship

Neither ”A” nor “B” follow each other



Dual Friendship

Both “A” and “B” follow each other

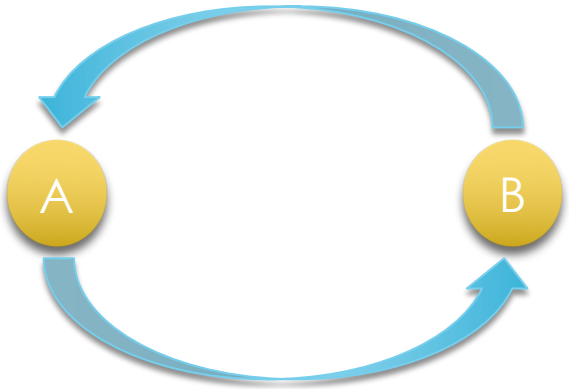
One-way Friendship

Either “A” or “B” follow the other

Relationship Categorize

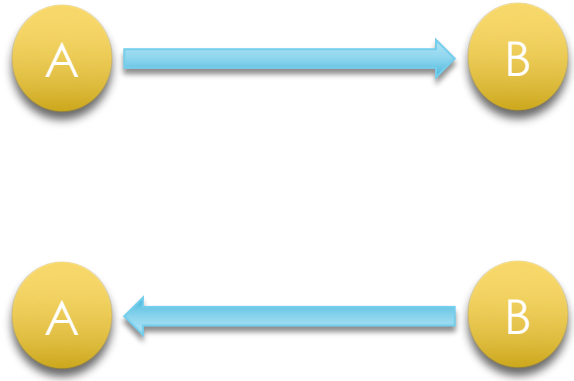
No Friendship

Not "A" nor "B" follow each other



Dual Friendship

Both "A" and "B" follow each other



One-way Friendship

Either "A" or "B" follow the other

Data collection & Annotation

Platform : Twitter

List of seed words : 65 words

Collected tweets : ~160K

Sampled tweets : 12K

Cleaned Labeled Tweets: 6.7K

- Choose the social media
- Design the seed words list
- Data processing:
  - Collect
  - Filter
  - Categorize
  - Clean/Pre-process
  - Sample
  - Annotate/Label

Data Collection & Annotation

- Gather a list of common and most used hostile words and use them as list of **seed word**
- Hatebase, an online lexicon of words that are used primarily in Hate Speech.
- Not every bad word in used in Hate Speech
- Online resources and Google list of profane words
- Examples:
  - ['f-word', 'n-word', 'c-word', 'b-word', 'r-word', idiot', 'queer', 'sex', and many more words]

Data Collection & Annotation

- Use the combination of different Twitter API functions
  - Stream function (statuses.filter) : Returns public statuses that match one or more provided filters.
  - Relationship function (friendships.show): Returns detailed information about the relationship between two arbitrary users.
- Filtered out unnecessary tweets:
  - Tweets which doesn't have the seed words
  - Non-English tweets
  - The starter tweets
  - The Self replied tweets

Data collection & Annotation

- Divide the data into 4 relationship categorizes
- Clean the data:
  - Removing the links and stopwords
  - Replaced the emojis
- Sample the data :
  - Picked 3K unique tweets from each relationship category
- Annotate the data:
  - Hostile/Non-hostile
  - Direct Hostility/ Non-direct Hostility

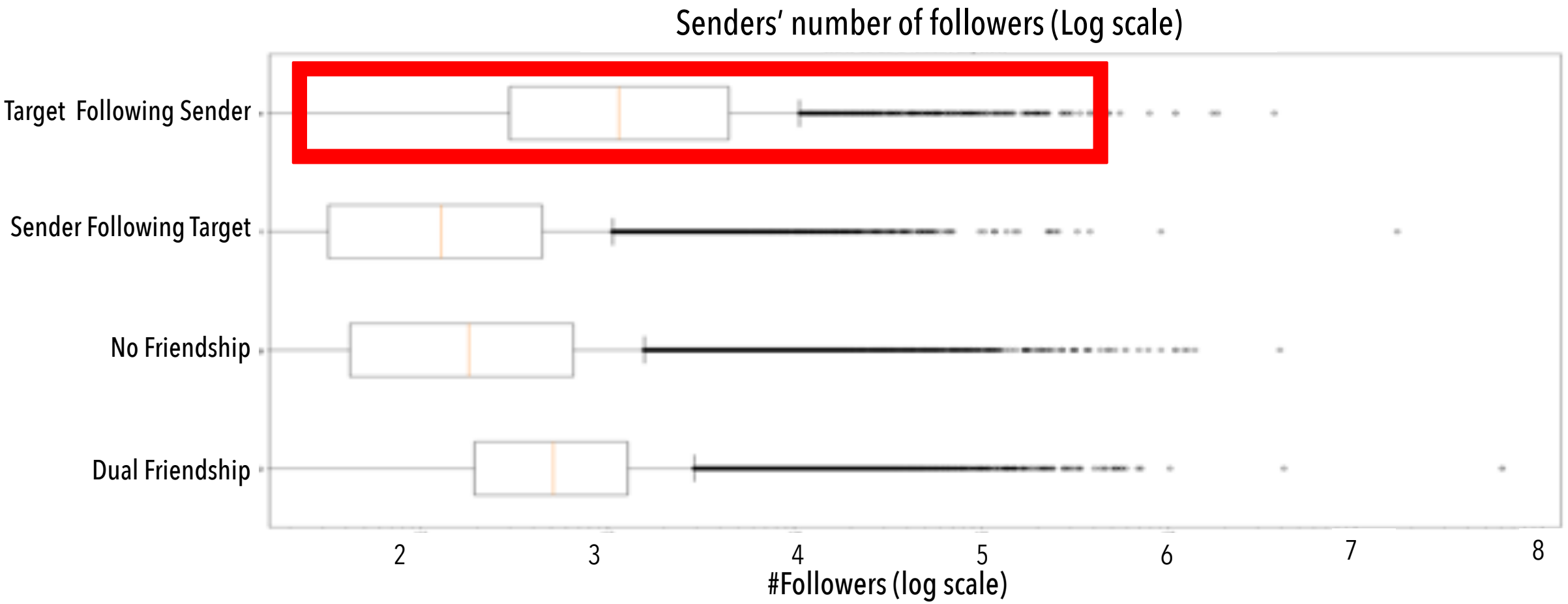
Data Analysis & Feature Extraction

- Update the tweets after 2 weeks
- Analyze tweets of each relationship category separately
- Extract the Features
  - Avg. #followers of Sender
  - Avg. #followers of Target
  - Deleted Tweets
  - #likes/retweets

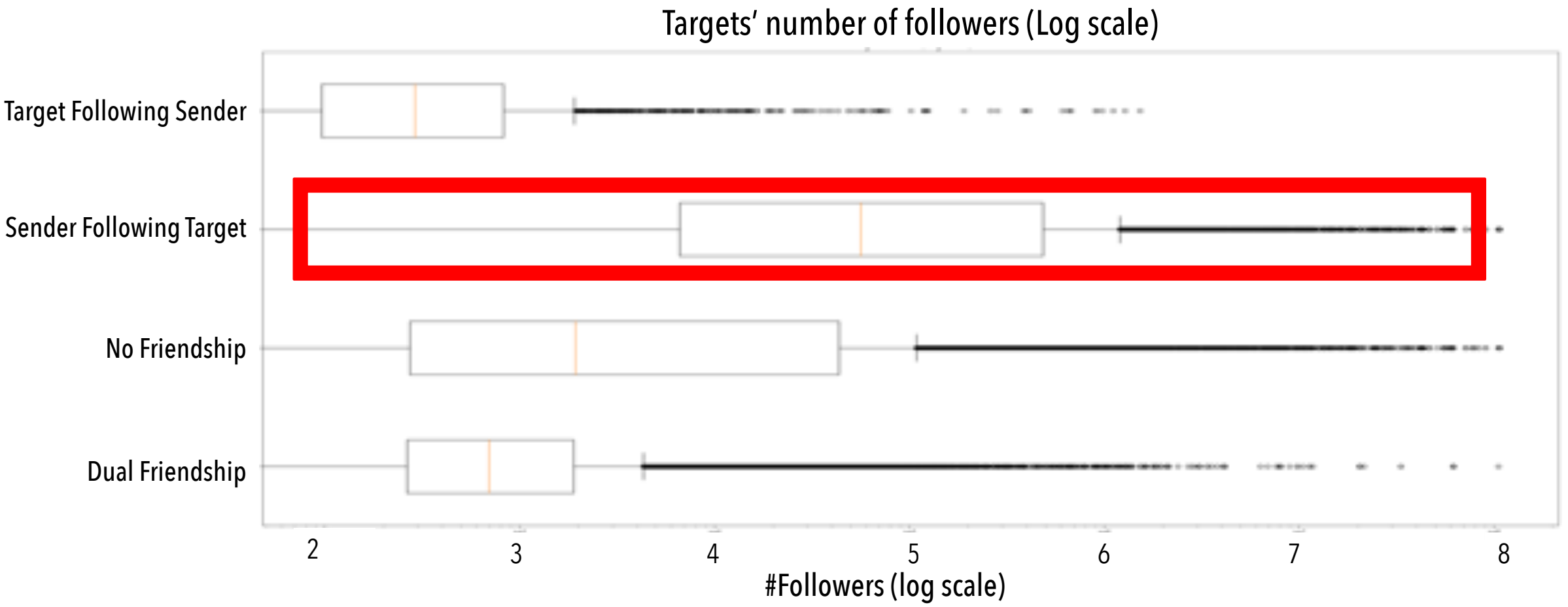
Deleted Tweets

Friendship Category	Size	Deleted Tweets	
		Quantity	Ratio
Dual Friendship	72,253	7,656	10.59%
No Friendship	50,548	8,917	17.64%
Source Follows the Target	31,697	4,089	12.90%
Target Follows the Source	2,429	357	14.70%

Senders' Followers

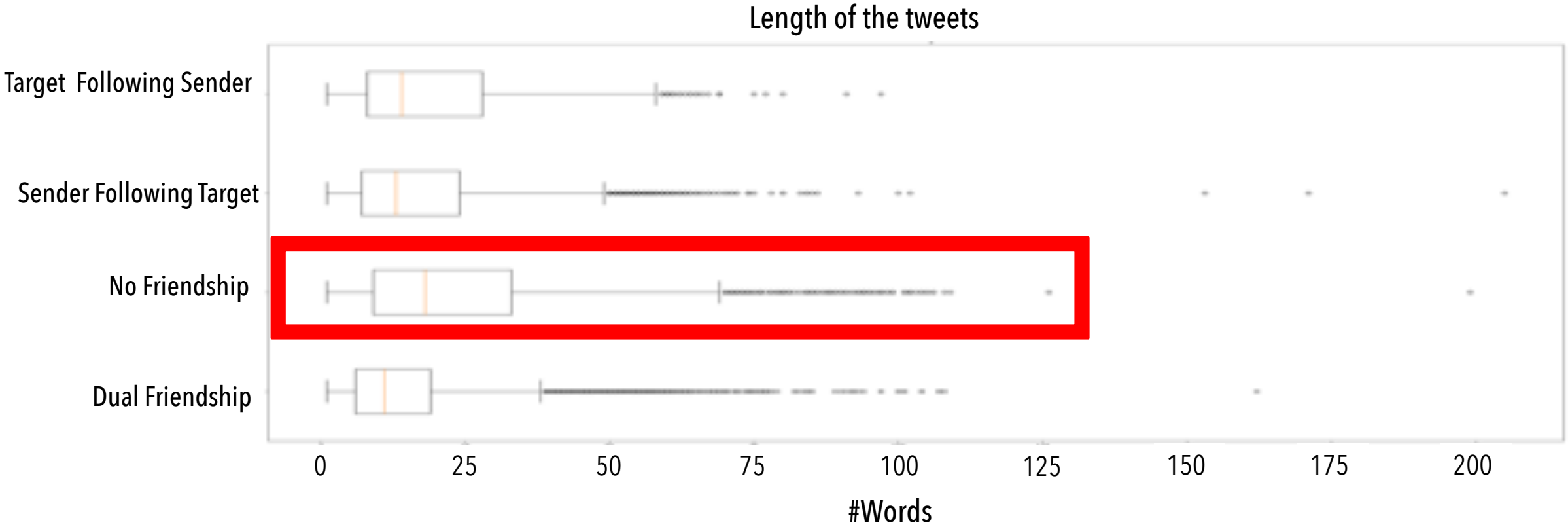


Targets' Followers



Tweets' Length

- There is a positive relationship between the length of the text and it's level of hostility.

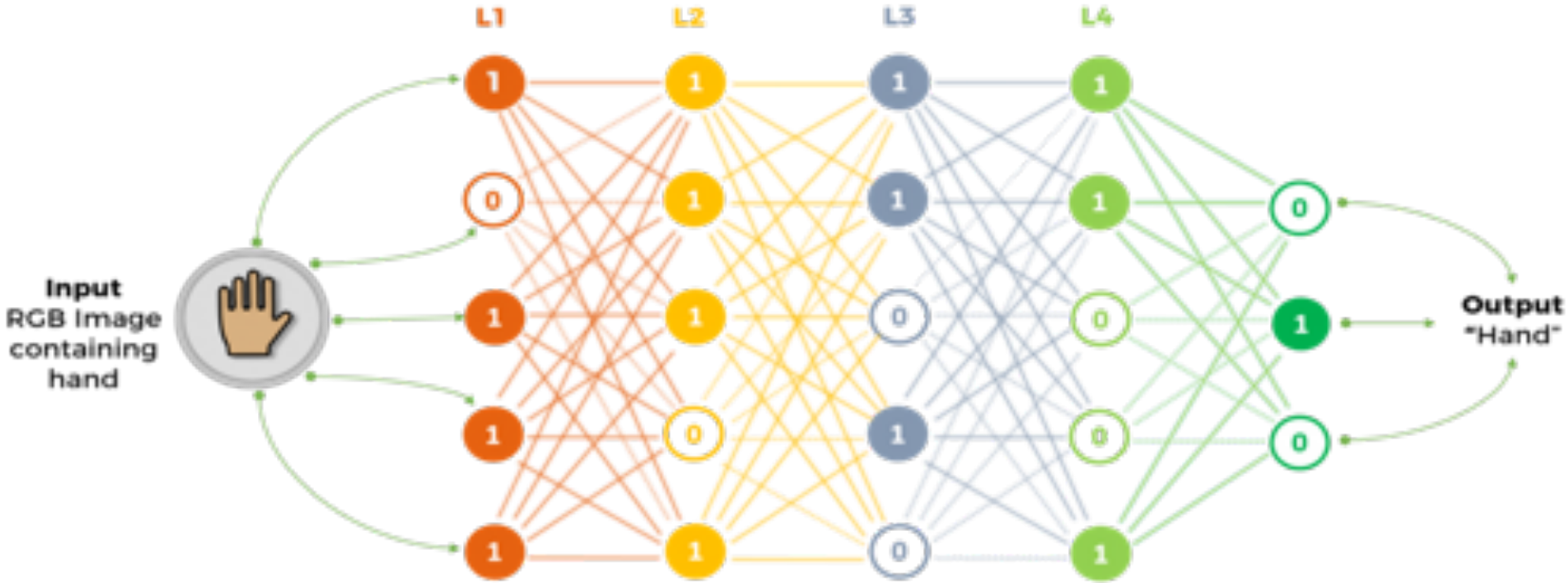


Classification Model

- Recurrent Neural Network (RNN)
- Long Short Term Memory (LSTM)
- Bi-directional LSTM
- Apply group features on classifiers

Neural network(NN)

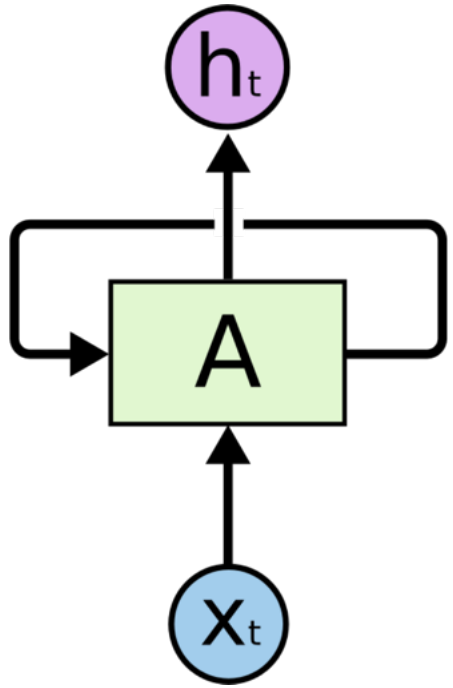
- A computer architecture in which a number of processors are interconnected in a manner suggestive of the connections between neurons in a human brain and which is able to learn by a process of trial and error



[1] https://www.manomotion.com/rise-machines-deep-learning/

Recurrent Neural Networks (RNN)

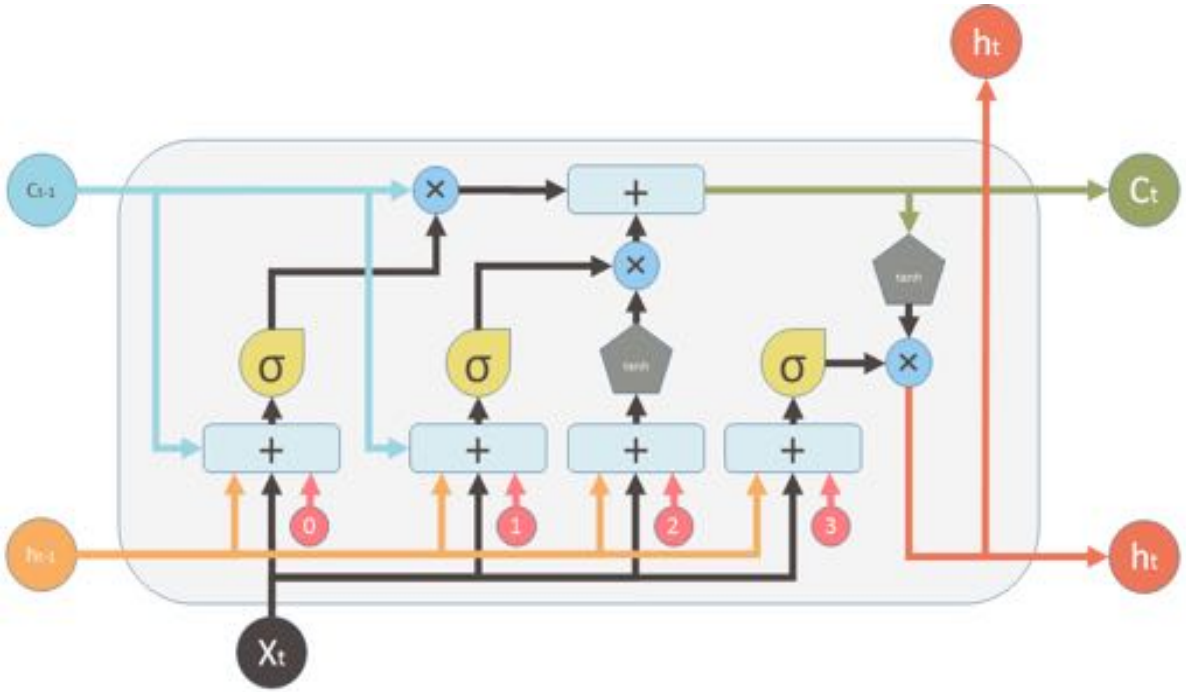
- A recurrent neural network can be thought of as multiple copies of the same network, each passing a message to a successor.
- The core reason that recurrent nets are more exciting is that they allow us to operate over sequences of vectors.



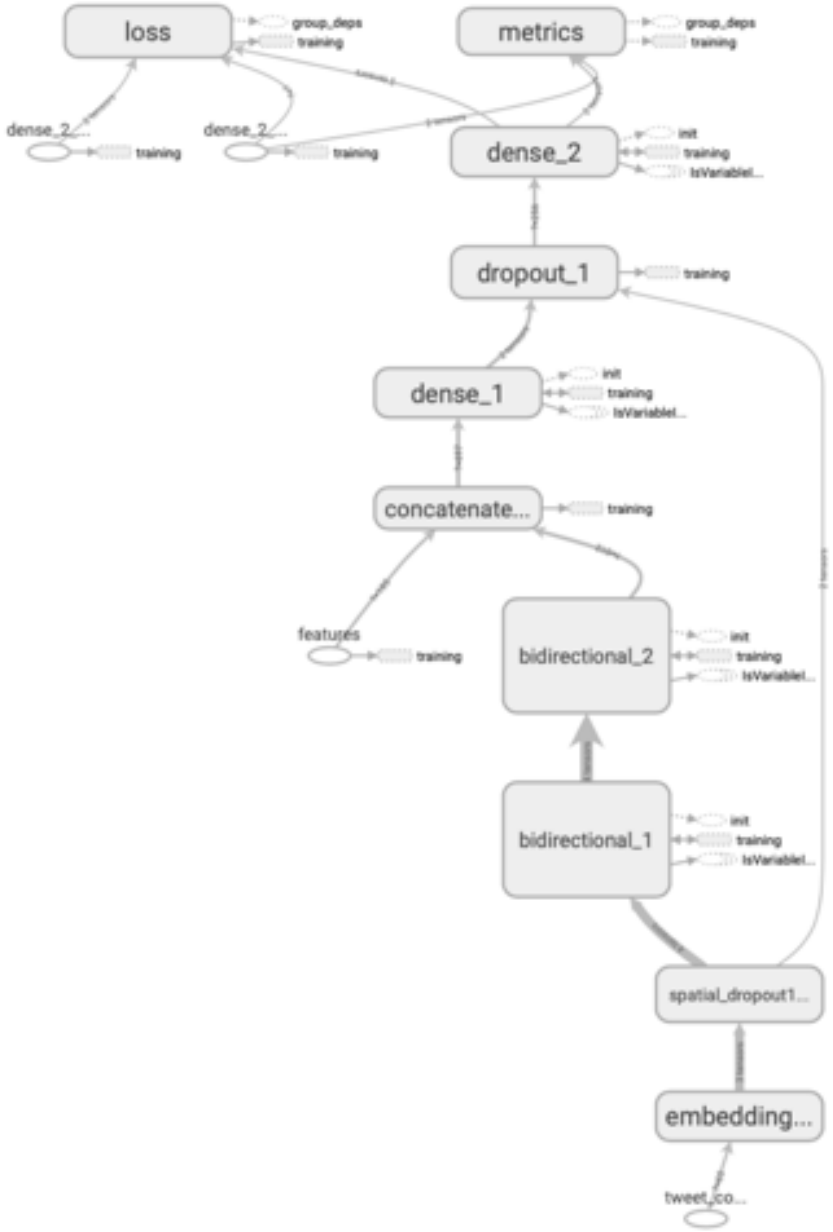
[1] https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Long Short Term Memory (LSTM)

- A special kind of RNN, capable of learning long-term dependencies.
- Designed to remember information for long periods of time
- **Bi-directional LSTM** connects two hidden layers of opposite directions for the same output.



[1] https://medium.com/mlreview/understanding-lstm-and-its-diagrams-37e2f46f1714



Classification Model

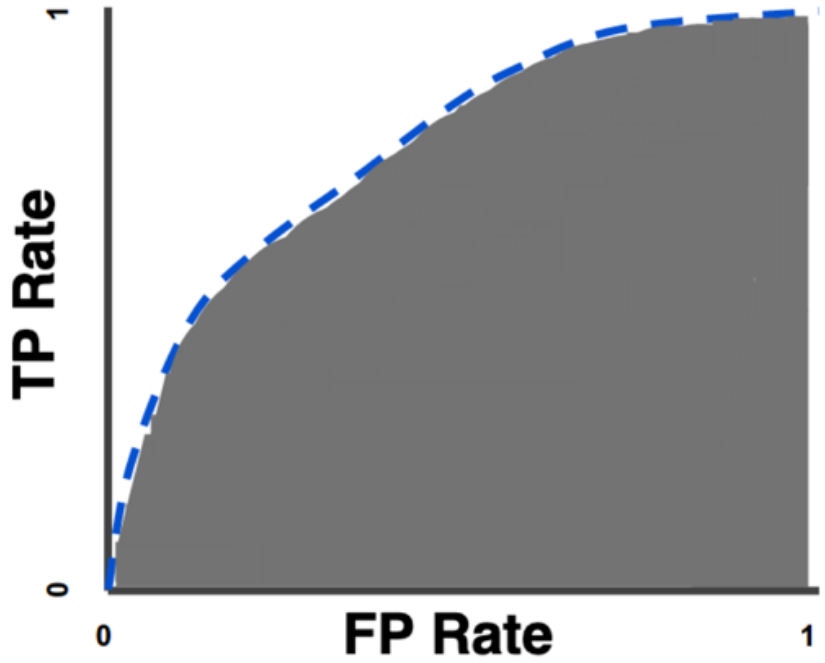
- Embedding layer(GloVe)
- Dropout
- Bidirectional LSTM
  - 256 cell LSTM
- Bidirectional LSTM
  - 256 cell LSTM
- Concatenation layer
- Dense (Relu)
- Dropout
- Dense (Softmax)

Evaluation

- Comparison Metrics
- Compare against a base model
- Analyze the results in details

Comparison Metrics

- Area Under the Curve (AUC):
  - The probability that the model ranks a random positive instance more highly than a random negative instance
- Key advantage of AUC:
  - More robust than accuracy precision, recall, and f1-measure in **class imbalanced situations**.
- Numerical classification metrics:
  - F1, precision, and recall



AUC (Area under the ROC Curve). [1]

[1] https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Base Model

- Receives the following features:
  - Full text of the tweet
  - Length of the tweet
- Trained with the same RNN
- Our proposed model only knows one more feature
  - Relationship between users

- Therefore:
  - Proposed Model = Base Model + Relationship Categories

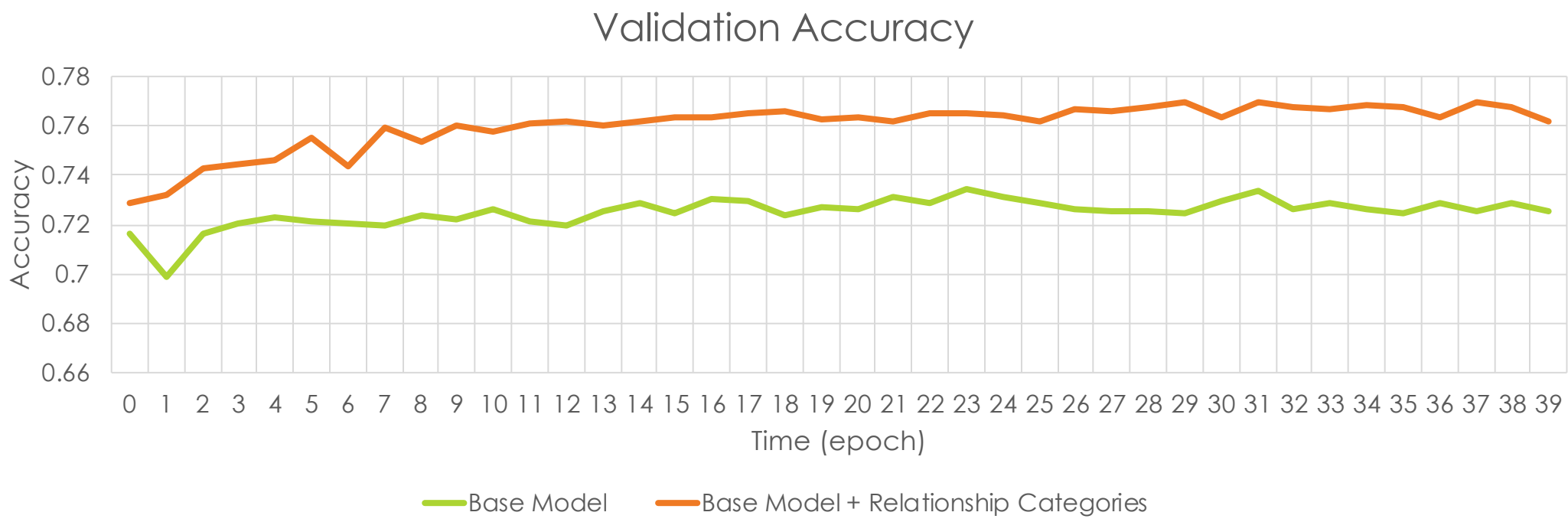
Evaluation – All tweets

Relationship category	No. of Tweets	Non-Hostile		Hostile	
		Quantity	Ratio	Quantity	Ratio
Dual Friendship	2,176	1,535	70.54%	641	29.46%
No Friendship	1,516	380	25.07%	1,136	74.93%
Sender Fallowing Target	1,610	623	38.70%	987	61.30%
Target Following Sender	1,469	611	41.59%	858	58.41%

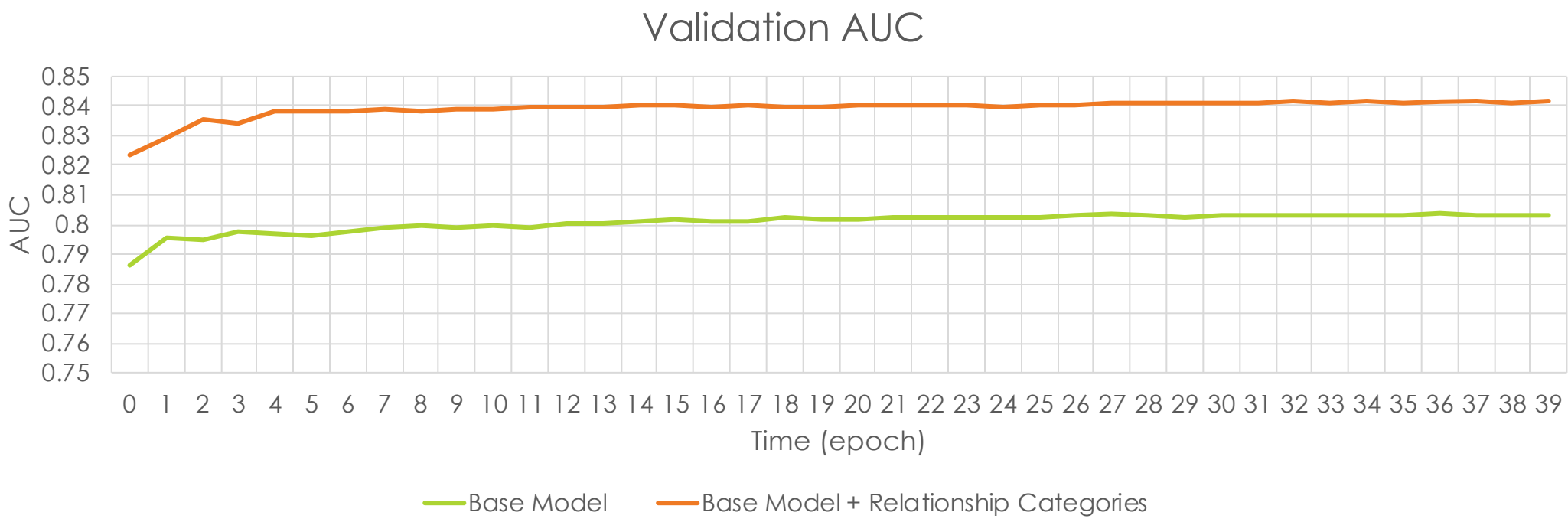
Evaluation – Hostile Tweets

Relationship category	No. of Hostile Tweets	Directed Hostility		Non-Directed Hostility	
		Quantity	Ratio	Quantity	Ratio
Dual Friendship	641	270	42.12%	371	57.88%
No Friendship	1,136	725	63.82%	411	36.18%
Sender Fallowing Target	987	599	60.69%	388	39.31%
Target Following Sender	858	507	59.09%	351	40.91%

Evaluation - Accuracy



Evaluation - AUC



F1, precision, and Recall

- Increase in numerical classification metrics

	Precision	Recall	F1
Base Model	0.7189	0.7121	0.7103
Our Model	0.7588	0.7532	0.7522

Future work

- Use a larger dataset (collect and label more data)
- Select smaller number of annotators
- Inferring more user attributes
  - Example: N-word in some countries
- Expand from sender/target discussion to a multi-people discussion

Conclusion

- Hostility in Social Media is elevating
- Previous works do not consider the relationship feature
- We sampled/labeled the data from Twitter and divide them into 4 main categories, each representing the relationship between the tweets participants
- Designed a LSTM model which receives the text, length of tweet, and relationship category as input and decide if the tweet is hostile of not
- After the evaluation model shows 5% improvement in AUC and 4% improvement in F1

[1] from Pinterest

[1] https://www.pinterest.com/pin/427630927109428916

Thank You!