Deep Learning Methods for the Detection of Cyberbullying in Social Media Messages.

Jack Leyland, Noura Al-Moubayed

3rd Year Project - Computer Science (with Year Abroad)

Description of background

Cyberbullying is a form of bullying that occurs over digital devices such as smartphones, computers and tablets[1]. Bullying in general is defined as 'unwanted, aggressive behaviour … that involves a real, or perceived power imbalance[2]. The behaviour is repeated, or has the potential to be repeated, over time.'. We can see how this could realise itself in the form of cyberbullying, often over text message, social media platforms, or email. It has been indicated that a huge 16% of students were bullied electronically in the 12 months prior to a 2015 survey alone.[1]

Deep Learning, put simply, is a subfield of Machine Learning, using algorithms inspired by the structure and function of the brain called artificial neural networks[3]. Generally, Deep Learning models process training data, make predictions, compare these with gold-standard data, and learn from errors to alter the model and improve the likelihood of correct future predictions.

I endeavour to experiment with a variety of Deep Learning methods, to learn the indicators of cyberbullying in social media messages and accurately detect if a given message is likely to be classed as cyberbullying or not.

Research question

Which Deep Learning architectures give the best results (on our dataset) for the classification of messages being cyber-bullying? Why?

Preliminary preparation

Reading:

- Using Machine Learning to Detect Cyberbullying, 2011.[4]
    - Uses Formspring data. Approximately 4000 posts. Used just the question text and the answer text. Hand-tagged as cyberbullying or not in 2011.
    - Manipulated the data. Best results were found by normalizing the number of 'bad' words with respect to the length of the post.
    - Used decision trees.
    - Reached 78.5% positive post identification on the development set with 8 repetitions of positive examples.

---

[1] 'What is Cyberbullying', Stop Bullying, 2018. https://www.stopbullying.gov/cyberbullying/what-is-it/index.html [19/06/18]

[2] 'What is Bullying', Stop Bullying, 2017. https://www.stopbullying.gov/what-is-bullying/index.html [19/06/18]

[3] Jason Brownlee, 'What is Deep Learning?', Machine Learning Mastery, 2016. https://machinelearningmastery.com/what-is-deep-learning/ [19/06/18]

[4] Kelly Reynolds, April Kontostathis, Lynne Edwards, 'Using Machine Learning to Detect Cyberbullying', Ursinus College, 2011. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.221.4788&rep=rep1&type=pdf [13/06/18]

    - Positive (bullying) examples composed only 10% of the dataset. Overall accuracy on an unseen test set reached 81.7% when positive examples were repeated in the dataset 8 times each, with positive identification being 67.4%.
    - They would like to extend this project in the future to consider profile information as well as just the text.

- Harassment Detection, a Benchmark in the #HackHarassment Dataset, 2016.[5]
    - Creates a new dataset, #HackHarrassment v.1.0 dataset, which improves the size and the quality of open source datasets for this task.
    - They hashed unigrams/bigrams/trigrams of the text, computed term frequency for each has value.
    - Got up to 73.3% accuracy. 80% precision, 71% recall with decision trees.
    - Attempted a deep learning approach. Reached 71% precision and 73% recall.
    - They were disappointed with these results, no improvement on Reynolds, 2011.

- Mean Birds: Detecting Aggression and Bullying on Twitter, 2017.[6]
    - Takes 1.6M tweets using the Twitter API, along with 30 features about the user profile and also who is in the user's social network.
    - Used 834 workers on CrowdFlower to hand-label the sessions as Normal, Bullying, Aggressive or Spammer. Here, a session is a small number of tweets coming from the same user. We classify the user as a bully, spammer etc. =
    - Took the tweets and profile data, extracted useful features, clean data (remove stop words, punctuation, make it lower case), then classify.
    - Achieved an impressive 89.9% precision, 91.7% recall, 91% accuracy with a random forest model.

- Cyberbullying Detection based on Text-Stream Classification, 2013.[7]
    - Uses an ensemble of one-class classifiers and session-based framework to tag large volumes of data from a stream of unlabelled text.
    - Uses the ensemble to extract only potential positive cyberbullying cases (containing swear words, for example), then train a network to classify these. Idea is that the network will more quickly learn to detect positive cyberbullying cases.
    - Out-performs the traditional fixed sliding-window approaches.
    - Data came from numerous sources (MySpace, Twitter, Kongregate), and lots of pre-processing was done such as removing most frequent words, hashtags etc.

- Efficient Estimation of Word Representations in Vector Space.[8]
- Deep Learning, Ian Goodfellow, Yoshua Bengio, Aaron Courville. [9]

---

[5] Alexei Bastidas, Edward Dixon, Chris Loo, John Ryan, 'Harassment Detection, a Benchmark in the #HackHarassment Dataset', Intel, 2016. https://pdfs.semanticscholar.org/c22d/fb9eff1c8c538d40a51609e7593cc9ded136.pdf [13/06/18]

[6] Despoina Chatzakou et al, 'Mean Birds: Detecting Aggression and Bullying on Twitter', Aristotle University of Thessaloniki, 2017. https://arxiv.org/pdf/1702.06877.pdf [13/06/18]

[7] Vinita Nahar, Xue Li, Chaoyi Pang, Yang Zhang, 'Cyberbullying Detection based on Text-Stream Classification', University of Queensland, 2013. http://crpit.com/confpapers/CRPITV146Nahar.pdf [13/06/18]

[8] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffry Dean, 'Efficient Estimation of Word Representations in Vector Space', 2013. https://arxiv.org/pdf/1301.3781.pdf [13/06/18]

[9] Ian Goodfellow, Yoshua Bengio, Aaron Courville, *Deep Learning book,* (The MIT Press, 2016) [13/06/18]

Datasets:

- The first dataset I found was taken from data from the Formspring site. Updated last in 2017. Reynolds, 2011, used a similar Formspring dataset but they created it themselves. This data set also has information of 'where' in the message the bullying occurred, which could be used for my advanced deliverable of predicting this with my own model on unseen examples – as well as the simple classification task. https://www.kaggle.com/swetaagrawal/formspring-data-for-cyberbullying-detection/data

- Next, I found a hand-annotated set of social media messages, used in the #HackHarassment paper. These were RAR files that I cannot open, the newest of which were created in 2016. http://chatcoder.com/DataDownload

- I then found an open source dataset of tweets from 2016, labelled with whether they were cyber-bullying or not on github. They achieve a good accuracy, but don't state what this figure is. This dataset seems extremely useful, however. https://raw.githubusercontent.com/varmichelle/Anti-Bully/master/datasets/new_data.csv

- The dataset from 'Mean Birds: Detecting Aggression and Bullying on Twitter', written in 2017, offers the tweet user id, the class of the user, and a set of 5-10 tweet ids for this user. Requires separate script utilising the Twitter API to extract the tweet itself. https://zenodo.org/record/1184178#.WyJ8PDMzayA

Potential problems

- Processing informal/badly spelled text.
- Dataset size (small number of positive examples). Prone to over-fitting.
- Size of examples, short text has limited context.
- Quality of data. Text might not be enough. Might want user information.

Project objectives

Minimum

1. Create and train a basic network that detects cyber-bullying in messages with an accuracy of greater than 50%. This is our minimum benchmark, as we would achieve 50% accuracy if we just randomly assigned a class to each instance.
2. Research existing work in this field and comment on approaches. Identify the current state of the art.
3. Evaluate the results with a variety of metrics. Look at the accuracy, precision, recall.
4. Justify the approach taken, suggest possible improvements.

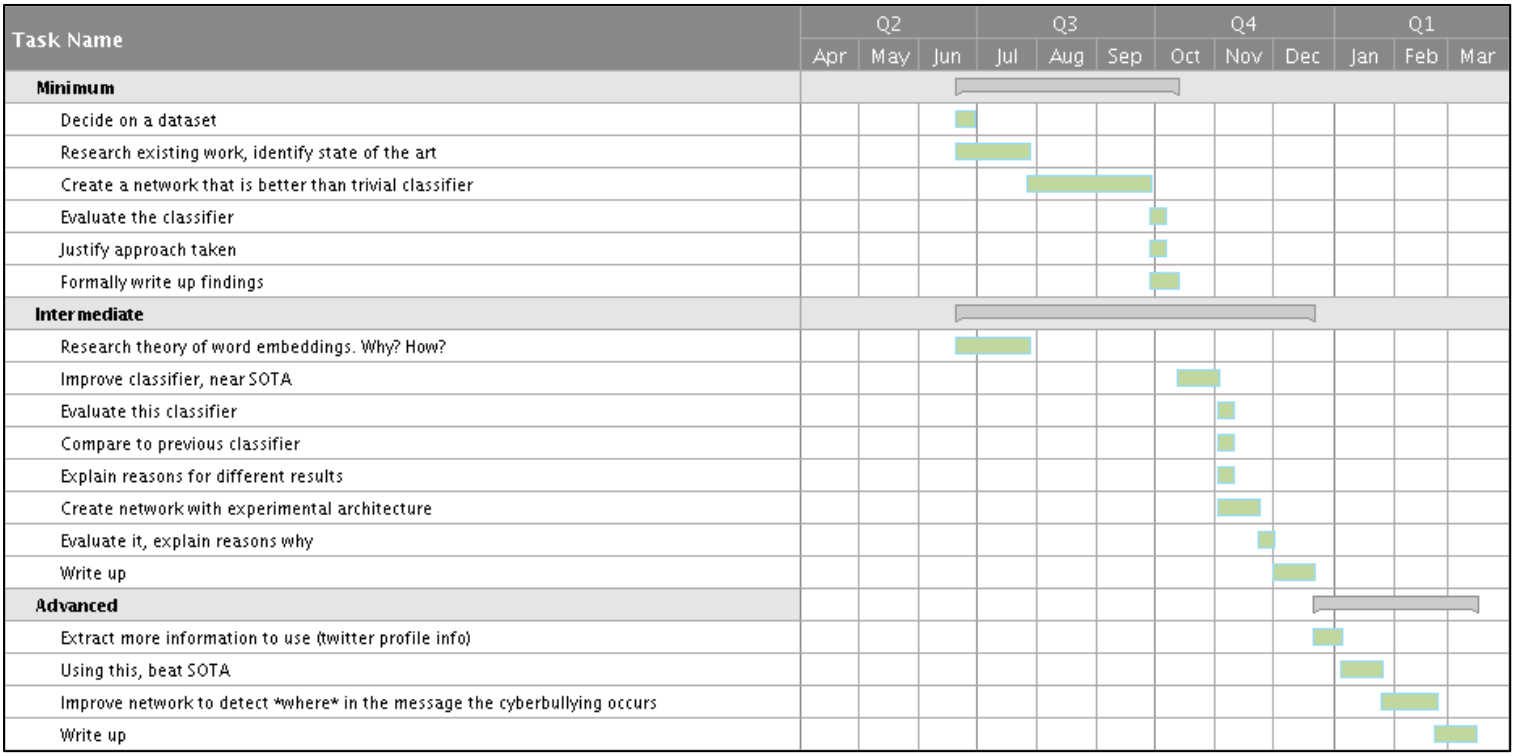
Intermediate (Minimum + …)

1. Create and train a network that **attempts to get close to** the state of the art results for this task with this dataset. Base this figure from approaches on the same dataset.
2. Explore calculating the word embeddings. How do we do this? Why do we do this?
3. Try to explain reasons why some approaches may be better than others.
4. Experiment with rare/unseen architectures of my creation. Explore their quality, try and explain reasons for them.

Advanced. (Intermediate + …)

1. Create and train a network that **beats** the state of the art results for this task with this dataset.
2. Extract more features such as user information. Could use Twitter API to do this for tweets/twitter profiles.
3. Additional output, calculate which part of the message is classed as bullying.

Gantt Chart



References

- Kelly Reynolds, April Kontostathis, Lynne Edwards, 'Using Machine Learning to Detect Cyberbullying', Ursinus College, 2011.
  http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.221.4788&rep=rep1&type=pdf [13/06/18]

- Alexei Bastidas, Edward Dixon, Chris Loo, John Ryan, 'Harassment Detection, a Benchmark in the #HackHarassment Dataset', Intel, 2016.
  https://pdfs.semanticscholar.org/c22d/fb9eff1c8c538d40a51609e7593cc9ded136.pdf [13/06/18]

- Despoina Chatzakou et al, 'Mean Birds: Detecting Aggression and Bullying on Twitter', Aristotle University of Thessaloniki, 2017.
  https://arxiv.org/pdf/1702.06877.pdf [13/06/18]

- Vinita Nahar, Xue Li, Chaoyi Pang, Yang Zhang, 'Cyberbullying Detection based on Text-Stream Classification', University of Queensland, 2013.
  http://crpit.com/confpapers/CRPITV146Nahar.pdf [13/06/18]

- Tomas Mikolov, Kai Chen, Greg Corrado, Jeffry Dean, 'Efficient Estimation of Word Representations in Vector Space', 2013.
  https://arxiv.org/pdf/1301.3781.pdf [13/06/18]

- 'Word2Vec', Tensorflow, Google, 2018.
  https://www.tensorflow.org/tutorials/word2vec [13/06/18]

- Ian Goodfellow, Yoshua Bengio, Aaron Courville, *Deep Learning book,* (The MIT Press, 2016).
  http://www.deeplearningbook.org/ [13/06/18]

- Jason Brownlee, 'What is Deep Learning?', Machine Learning Mastery, 2016.
  https://machinelearningmastery.com/what-is-deep-learning/ [19/06/18]

- 'What is Bullying', Stop Bullying, 2017
  https://www.stopbullying.gov/what-is-bullying/index.html [19/06/18]

- 'What is Cyberbullying', Stop Bullying, 2018
  https://www.stopbullying.gov/cyberbullying/what-is-it/index.html [19/06/18]