Aggression Identification or Depression Detection using Generative Adversarial Networks

Andriy Drozdyuk and Norbert Eke

1 Problem Description

Depression detection and aggression identification are two important classification tasks in natural language processing. Nowadays more and more social media data is filled with offensive language, hate speech, cyber bullying, which endangers the cyber-safety of both children and adults online. Automated aggressive language detection would lighten the workload on website and chat-room moderators whose job it is to maintain a safe communication and interaction in a cyber-society. Depression detection is a more serious problem that could be addressed using automated text classification systems. There is a need for a system that detects users at risk of depression.

In our project we are looking to address either the problem of depression detection or aggression detection. Using the publicly available data sets depression detection would be a two class classification problem (depressive and non-depressive text), while aggression detection would be a three class classification task (overtly aggressive, covertly aggressive and non-aggressive texts). Ideally we would like to work on a two class classification approach, which is depression detection, but the decision to pick one task over the other will come down to whether or not we can gain access to the entire depression detection data set. This data set contains annotated tweet ids, but not the tweets themselves, thus we need to crawl/download each tweet by its tweet id.

2 Initial Ideas

We had three initial approaches to either depression or aggression detection, but ultimately decided to pursue the first one:

1. *Design an approach to combat unbalanced classification tasks using Generative Adversarial Networks. This task can be thought of as customized anomaly detection.*
2. Design a two stage classification system for aggression detection (the only task containing 3 classes) by first having a classifier differentiate between non-aggressive and everything else, then have a second classifier take what is left and classify as openly-aggressive vs. covertly aggressive.
3. The training and testing data for aggression detection come from different social networking platforms. There was an initial idea about using the word embeddings and labels of the training data to learn better word embeddings for the testing data.

One of the winning entries to aggression identification [2] attribute 5% accuracy increase to data augmentation. Generative Adversarial Networks (GAN) are known for their ability to generate more data. Since both domains of aggression identification and depression detection contain very unbalanced data sets (i.e. there are many more examples of non-depression tweets then ones with depression), this lead us to explore the notion of using GANs.

GAN is a deep generative model. It is built on the framework of *adverserial nets*, where a generative model is put in competition with a descriminative model. The discriminative model learns to distinguish between samples from model distribution and data distribution. In the special case of generative model creating samples by passing noise through a multi-layer perceptron and discriminative model also being a multi-layer perceptron, we call the result the *adverserial nets*.

Let z represent the noise and the p_z the distribution over z. When z is acted upon by the generator it produces x in the data space, with a corresponding distribution p_g. Discriminator takes x and produces a single scalar.

If $G(z;\theta_g)$ is a multi-layer perceptron with parameters θ_g, it represents the generator that maps from the noise to the data space. Then if $D(x;\theta_d)$ is another multilayer perceptron with parameters θ_d, it represents the discriminator that maps from the data space to a scalar.

We train D to assign correct label to data and to samples generated by G. Correct in this case means $D(x;\theta_d)$ having high probability if the x came from data, and low probability of it came from p_g. At the same time, we train G to minimize $\log(1 - D(G(z;\theta_g);\theta_d))$.

Our hypothesis is that we can apply GAN to classify text by first training the generative model to produce instances of one class, and then using the discriminative model to differentiate between normal and anomalous data. Anomalous data in this case would be instances of either aggressive or depressive text. Our contribution is in using *text patches* to train the model, instead of traditional approaches. This approach follows that of [7], who used it on images. We hope that by using GANs in conjunction with *text patches* we will be able to achieve better accuracy on classification of non-symmetric data, such as often present in depression or aggression classification.

3 Previous work

In 2014 Goodfellow et. al. [5] proposed the idea of Generative adverserial nets, which consisted of two models: the generative and discriminative. The generative model can *generate* inputs to the discriminative model. The discriminative model estimates the probability that the input came from the real data rather than the generative model. The generative model's goal is to maximize the probability of discriminative model making a mistake.

In [4], authors proposed a GAN model that detects anomalies in Air Surveillance data to support real time military decisions making. In [1], authors proposed FakeGAN, an augmentation to the GAN model which is effective at detecting deceptive reviews. In [3], authors performed image anomaly detection with Generative Adversarial Networks. In [7], authors proposed an anomaly detection approach using GANs for medical images based on image patches randomly being located and extracted from the original image.

The task of aggression identification has been posed as a challenge in the First Workshop on Trolling, Aggression and Cyberbullying [6]. The competitors were tasked to classify the dataset as either *Overtly Aggressive*, *Covertly Aggressive*, and *Non-aggressive*. The winning entry obtained a weighted F-score of 0.64. One of the best entries [2] used LSTM, CNN and an ensemble of the two. Authors reported a 5% accuracy increase due to data augmentation.

References

[1] Hojjat Aghakhani et al. "Detecting Deceptive Reviews Using Generative Adversarial Networks". In: *2018 IEEE Security and Privacy Workshops (SPW)* (May 2018). DOI: 10.1109/spw.2018.00022. URL: http://dx.doi.org/10.1109/SPW.2018.00022.

[2] Segun Taofeek Aroyehun and Alexander Gelbukh. "Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling". In: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. 2018, pp. 90–97.

[3] Lucas Deecke et al. *Anomaly Detection with Generative Adversarial Networks*. 2018. URL: https://openreview.net/forum?id=S1EfylZ0Z.

[4] Fahrettin Gökgöz. "Anomaly Detection using GANs in OpenSky Network". In: *NATO Science and Technology Organization: Big Data and Artificial Intelligence for Military Decision Making*. May 2018. URL: https://www.sto.nato.int/publications/STO%5C%20Meeting%5C%20Proceedings/STO-MP-IST-160/MP-IST-160-W2-2.pdf.

[5] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[6] Ritesh Kumar et al. "Benchmarking Aggression Identification in Social Media". In: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. 2018, pp. 1–11.

[7] Thomas Schlegl et al. "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery". In: *International Conference on Information Processing in Medical Imaging*. Springer. 2017, pp. 146–157.