**LITERATURE SURVEY**

# “An optimization-based approach for sequential label aggregation,” in Findings of the Association for Computational Linguistics:”,

# AUTHORS: N. Sabetpour, A. Kulkarni, and Q. Li, “OptSLA:

# The need for the annotated training dataset on which data-hungry machine learning algorithms feed has increased dramatically with advanced acclaim of machine learning applications. To annotate the data, people with domain expertise are needed, but they are seldom available and expensive to hire. This has lead to the thriving of crowdsourcing platforms such as Amazon Mechanical Turk (AMT). However, the annotations provided by one worker cannot be used directly to train the model due to the lack of expertise. Existing literature in annotation aggregation focuses on binary and multi-choice problems. In contrast, little work has been done on complex tasks such as sequence labeling with imbalanced classes, a ubiquitous task in Natural Language Processing (NLP), and Bio-Informatics. We propose OptSLA, an Optimization-based Sequential Label Aggregation method, that jointly considers the characteristics of sequential labeling tasks, workers reliabilities, and advanced deep learning techniques to conquer the challenge. We evaluate our model on crowdsourced data for named entity recognition task. Our results show that the proposed OptSLA outperforms the state-of-the-art aggregation methods.

# 2) “Assessing gan-based approaches for generative modeling of crime text reports,”

# AUTHORS: S. Khorshidi, G. Mohler, and J. G. Carter

# Analysis and modeling of crime text report data has important applications, including refinement of crime classifications, clustering of documents, and feature extraction for spatio-temporal forecasts. Having better neural network representations of crime text data may facilitate all of these tasks. This paper evaluates the ability of generative adversarial network models to represent crime text data and generate realistic crime reports. We compare four state of the art GAN algorithms in terms of quantitative metrics such as coherence, embedding similarity, negative log-likelihood, and qualitatively based on inspection of generated text. We discuss current challenges with crime text representation and directions for future research.

# 3 )“Satellitener: An effective named entity recognition model for

# the satellite domain,"

# AUTHORS: O. Jafari, P. Nagarkar, B. Thatte, and C. Ingram.

Nowadays, large amounts of data is generated daily. Textual data is generated by news articles, social media such as Twitter, Wikipedia, etc. Managing these large data and extracting useful information from them is an important task that can be achieved using Natural Language Processing (NLP). NLP is an artificial intelligence domain dedicated to processing and analyzing human languages. NLP includes many subdomains such as Named Entity Recognition (NER), Entity Linking, Sentiment Analysis, Text Summarization, Topic Modeling, and Speech Processing.

**4 “A survey on machine learning techniques for auto labeling of video, audio, and text data,”**

**AUTHORS: S. Zhang, O. Jafari, and P. Nagarkar**

Machine learning has been utilized to perform tasks in many different domains such as classification, object detection, image segmentation and natural language analysis. Data labeling has always been one of the most important tasks in machine learning. However, labeling large amounts of data increases the monetary cost in machine learning. As a result, researchers started to focus on reducing data annotation and labeling costs. Transfer learning was designed and widely used as an efficient approach that can reasonably reduce the negative impact of limited data, which in turn, reduces the data preparation cost. Even transferring previous knowledge from a source domain reduces the amount of data needed in a target domain. However, large amounts of annotated data are still demanded to build robust models and improve the prediction accuracy of the model. Therefore, researchers started to pay more attention on auto annotation and labeling. In this survey paper, we provide a review of previous techniques that focuses on optimized data annotation and labeling for video, audio, and text data.

5) **“Cyberbullying detectionusing time series modeling**

**AUTHORS**: **NektariaPotha and ManolisMaragoudakis.**

Cyber bullying is a new phenomenon resulting from the advance of new communication technologies including the Internet, cell phones and Personal Digital Assistants. It is a challenging bullying problem occurring in a new territory. Online bullying can be particularly damaging and upsetting because it's usually anonymous or hard to trace. In this paper, the proposed method is utilizing a dataset of real world conversations (i.e. Pairs of questions and answers between cyber predator and the victim), in which each predator question is manually annotated in terms of severity using a numeric label. We approach the issue as a sequential data modelling approach, in which the predator's questions are formulated using a Singular Value Decomposition representation. The motivation of this procedure is to study the accuracy of predicting the level of cyber bullying attack using classification methods and also to examine potential patterns between the lingustic style of each predator. More specifically, unlike previous approaches that consider a fixed window of a cyber-predator's questions within a dialogue, we exploit the whole question set and model it as a signal, whose magnitude depends on the degree of bullying content. Using feature weighting and dimensionality reduction techniques, each signal is straightforwardly parsed by a neural network that forecasts the level of insult within a question given a window between two and three previous questions. Throughout the time series modeling experiments, an interesting discovery was made. By applying SVD on the time series data and taking into account the second dimension (since the first is usually modeling trivial dependencies between instances and attributes) we observed that its plot was very similar to the plot of the class attribute. By applying a Dynamic Time Warping algorithm, the similarity of the aforementioned signals was proved to exist, providing an immediate indicator for the severity of cyber bullying within a give...