CIS 530 Final Project Proposals

Paul Zuo, Rani Iyer, Anosha Minai, Claire Wang, Graham Mosely, Sam Akhavan

Project 1: Identifying Tweets by Russian Trolls

1. **Problem Definition**

"As part of the House Intelligence Committee investigation into how Russia may have influenced the 2016 US Election, Twitter released the screen names of almost 3000 Twitter accounts believed to be connected to Russia's Internet Research Agency, a company known for operating social media troll accounts. "

- Kaggle

Our project would be using various methodology we learned this semester to extract features and attributes from given data, and to train a classifier or neural network that is able to most accurately predict whether a given tweet is a troll tweet or not.

2. **References**

Exposing Paid Opinion Manipulation Trolls
Proceedings of Recent Advances in Natural Language Processing
, pages 443–450,
Hissar, Bulgaria, Sep 7–9 2015.
https://www.aclweb.org/anthology/R/R15/R15-1058.pdf
Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying

http://paginaspersonales.deusto.es/isantos/papers/2013/2013-Patxi-Cyberbullying.pdf

Comment Abuse Classification with Deep Learning
https://web.stanford.edu/class/cs224n/reports/2762092.pdf
NLP Hacks: Trolling the trolls using Natural Language Processing, Intercom and AWS Lambda
http://blog.aylien.com/nlp-hacks-trolling-the-trolls-using-natural-language-processing-intercom-and-aws-lambda/
Mean Birds: Detecting Aggression and Bullying on Twitter
https://arxiv.org/pdf/1702.06877.pdf
Identifying political bot/troll social media activity using machine learning
https://medium.com/@conspirator0/identifying-political-bot-troll-social-media-activity-using-machine-learning-20dcd56e961a
Modeling Trolling in Social Media Conversations
https://www.utdallas.edu/~lxm111830/content/trolling.pdf#!
Troll Detection with Scikit-Learn
http://blog.kaggle.com/2012/09/26/impermium-andreas-blog/
Accurately Detecting Trolls in Slashdot Zoo via Decluttering

https://cs.umd.edu/~srijan/pubs/trolls-asonam14.pdf

3. Evaluation Metrics

Since it is a binary classification problem, we use accuracy, precision, recall and f1-score to evaluate the performance of classifier on both development and testing data.

4. Data Description

"This dataset contains two CSV files. tweets.csv includes details on individual tweets, while users.csv includes details on individual accounts."

- Kaggle

The dataset we are using are the 3000 twitter accounts and 200,000 tweets deleted by Twitter: https://www.kaggle.com/vikasg/russian-troll-tweets. We plan to separate the data into training, development and testing.

Project 2:

RumourEval: Determining rumour veracity and support for rumours

1. Problem Description

The Internet gave rise to the Information Age, dramatically lowering the barrier for receiving and sending information. The truth of this information, however, has never been guaranteed. Over the past several years, "fake news" disseminating on the Internet (originating from the ignorant and, more troublingly, the malicious) has had a serious impact on elections and political movements. But the Internet is too enormous to be fact-checked by journalists. How can we develop automated systems to evaluate the veracity of rumors?

In this task, we attempt to evaluate rumors and label them as true or false, with an associated confidence metric to capture the ambiguity involved in the problem (which may result in a label of "unverified"). For each rumor, we have a list of responses (the data is from Twitter). For the first subtask of this problem, we will label each response to each rumor as either: supporting, denying, questioning, or commenting. For the second subtask (which will be our main focus of the project) we will evaluate whether rumors are true or false. We will use the types of the responses as a feature, as well as utilize the external information provided in the

2. References

Many papers submitted to the SemEval 2017 conference have results and methods that are useful for establishing a baseline of features and evaluation metrics. We found "Determining Rumour and Veracity Support for Rumours on Twitter" by Omar Enayet and Samhaa R. El-Beltagy, http://www.aclweb.org/anthology/S17-2082. Another submitted paper," Detecting Stance towards Rumours with Topic Independent Features" by Hareesh Bahuleyan and Olga Vechtomov, will also give us more options about what features to extract, http://www.aclweb.org/anthology/S17-2080. We can also learn from the definitions of rumours used for the task, and the ways that rumours have been labeled in the training data from a paper called "Analysing How People Orient to and Spread

Rumours in Social Media by Looking at Conversational Threads" by A. Zubaiga et al : http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0150989. The paper describing the task in depth itself is also useful, http://www.derczynski.com/sheffield/papers/rumoureval-task.pdf, as are the referenced papers. We can also refer to Chapters 4 and 8 in the course textbook (Jurafsky et al) for more information about language model-based classification tasks.

3. **Evaluation Metrics**

Our problem has to two tasks: a multi-class classification task (categorizing responses to rumors) and a binary classification task (is rumor true or false). For the first task, we will assign f-scores to each category and then calculate an overall f-score for all categories. This approach is similar to how we evaluated our NER system in HW7. For the second task, we will follow the evaluation approach specified by SemEval, which is to calculate the microaccuracy, aka the ratio of instances for which a correct prediction is made.

4. **Data Description**

We will use the data provided by the SemEval shared task. The data are sets of tweets for eight different incidences / topics that would be likely to generate rumors. For each incident, there are about 25 source tweets (the rumors) and then some number of response tweets to each source tweet. We also have a dump of Wikipedia articles related to the incidences that are from before the eventual verification / denial of the rumors, which will simulate the use of external news sources (from the time of the rumor) in helping to verify rumors.

Project 3: Detecting Clickbait Titles for News Articles

1. **Problem Definition**

The issue of misinformation in the media has been a central concern in American politics recently. Nowadays, we often associate this with "fake news", but misinformation, particularly through the issue of misleading headlines, is not specific to news. It can include a wide variety of content, from fabricated content to satire, each of which would require a different background for analysis. With our research goal, we hope to investigate the problem of headline incompatibility.

Many headlines today are manipulated and twisted. For instance, the headline "Air pollution now leading cause of lung cancer" would be considered a clickbait headline because the evidence points to the fact that "outdoor air pollution is not only a major risk to health in general, but also a leading environmental cause of cancer deaths." This would be considered clickbait because, for one, omitting "environmental" from the headline largely generalizes the claim. Secondly, omitting the indefinite determiner 'a' may lead some readers to infer air pollution to be "the" leading cause of lung cancer.

Detecting clickbait headlines also comes with strong business implications. Many companies today are trying to determine whether or not an advertisement is a clickbait advertisement.

Clickbait ads are those ads with a very catchy headline that induce people to click on them. But the destination page after clicking has little to do with the headline. Companies want to be able to avoid these clickbait ads because though they will lead to a high initial clickthrough rate, the clickthrough rate will drastically decrease in the long term, hurting the company's marketing reputation and strategy.

2. References

http://www.aclweb.org/anthology/W17-4210
https://www.ijcai.org/proceedings/2017/0583.pdf
https://arxiv.org/pdf/1610.09786.pdf

3. Evaluation Metrics

We can go with two different evaluation criteria for this problem. One is using a simple accuracy score for the predicted labels (F1, accuracy rate, recall, sensitivity), with the labels being related (headline and body of text are related) or unrelated (headline and body of text are unrelated).

A further evaluation metric can be derived for more specific labels, like whether the headline and body of text has a relationship that agrees, disagrees or discusses. This would be a more nuanced dive into the simpler binary labels as above. We could use different multi-class classification evaluation metrics here.

Additional things to consider might include ROC curve optimization, looking at the different tradeoffs between sensitivity and specificity. What is the cost of classifying a non-clickbait headline as clickbait versus vice-versa? These are important questions when considering a binary classifier. Similarly, for a multi-class classifier, we may want to consider different weighted penalties for misclassifying agreement, disagreement of discussment.

4. Data Description

In his paper, Chakraborty crowdsourced labels for 7,500 clickbait articles and 7,500 non-clickbait articles. The data that he used for his research has been made publicly available. Additionally, the Fake News Challenge has a dataset that's derived from the Emergent Dataset created by Craig Silverman. This training data has headline, body, label tuples and the testing set has the same format but without the labels.