Stance Detection for Fake News Challenge Dataset using Deep Learning

Project for DS5220 by -
Shishir Kurhade and AANB Trihatmaja

Introduction

- 64% of US adults said that fake news has caused a great deal of confusion about the basic facts of current issues and events [1]
- To face this issue, we want to automatically classify the news into four categories (stances): unrelated, discuss, agrees, disagrees
- This problem is based on Fake News Challenge [2]

[1] Michael Barthel, Amy Mitchell, and Jesse Holcomb. Many americans believe fake news is sowing confusion, Dec 2016.
[2] http://www.fakenewschallenge.org/

Introduction

A reasoning for these labels is as follows:

1. Agrees: The body text agrees with the headline.
2. Disagrees: The body text disagrees with the headline.
3. Discusses: The body text discuss the same topic as the headline, but does not take a position
4. Unrelated: The body text discusses a different topic than the headline

Data Set Overview

There are two csv files:

1. train bodies.csv : contains the body text of articles (the articleBody column) with corresponding IDs (Body ID)
2. train stances.csv : contains the labeled stances (the Stance column) for pairs of article headlines (Headline) and article bodies (Body ID, referring to entries in train_bodies.csv)

Rows	49972
Unrelated	0.73131
Discuss	0.17828
Agree	0.0736012
Disagree	0.0168094

Dataset distribution

Machine Learning Methods

We plan on using Recurrent Neural Nets to solve this text classification problem.

1. To counter the vanishing gradient problem there are two variations of RNNs viz. LSTM and GRUs
2. GRU unit controls information flow across units like LSTM but without using memory unit and exposes the entire hidden state
3. GRUs are computationally more efficient and structurally less complex

Machine Learning Methods

Data Pre-processing

- Convert text from the corpus to tokens using **nltk** package
- Map text to corresponding vectorized forms using **GloVe** representations
- Normalize the case, handling the punctuation and non-alphabetic symbols

Modeling

- RNN GRU will be use
- Logistic regression is used as a baseline

Evaluation

- Use scoring system provided by Fake News Challenge
- Unrelated / Related weighted as 25%
- Agree / disagree / discuss weighted as 75%

Thank you