Fake News Stance Detection

Xiaowei Wu, Sizhu Cheng, Zixian Chai
{wux, scheng72, zchai14}@stanford.edu
All authors contributed equally

Problem

Fake news is deliberate misinformation fabricated with intention of deception, misleading, grabbing attention or even financial and political gain. Recent development of machine learning provides a possible solution to automate this process. However, accurately and repeatedly identifying fake news is still proven difficult due to the complex nature of human language. With the popularity of online media and detrimental effect of fake news on many aspects of our society, developing a reliable machine learning model for fake news identification becomes very important.

Data

The data set provided Fake News Challenge[1] will be used for this task. It consists of 49972 instances, each with a headline and a body text as input and the stance as output.

Example of an instance:

Body ID: 4

Headline: *It Begins: HazMat-Wearing Passenger Spotted At Airport*

Body: *Last week we hinted at what was to come as Ebola fears spread across America. Today, we get confirmation. As The Daily Caller reports, one passenger at Dulles International Airport outside Washington, D.C. is apparently not taking any chances.*

Stance: discuss

Features

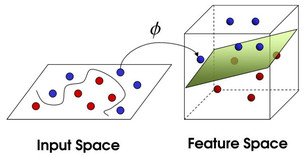
In addition to features such as the number of overlapped words, number of overlapped n-grams and number of negative words in the headlines, the following features are also implemented:

- **Similarity features**: The text is converted to a sparse vector to indicate the frequency of each word. The cosine of the angle between these two vectors is calculated as an indication of similarity.
- **BOW(bag-of-word) features:** Each document is defined as an N-dimensional vector $\{w_1{:}c_1, w_2{:}c_2 ... w_n{:}c_n\}$ in the vocabulary space with each word given a specific weight value c. The BOW vector for the headline-body pair is defined by overlapped words with binary weight.
- **Word Sentiment features**: WordNet was used to figure out all the verb synonyms of the words inside the provided "refuting_word" list, which indicates the sentiment.
- **Polarity features**: The NRC Word-Emotion Association Lexicon[3] is used to obtain a list of negative emotion associated English words. The polarity of a body is determined by the number of negative words

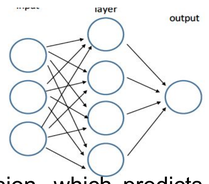
Models

Using scikit learn[3], these models are implemented to learn from the training data using k-fold (k=10) cross-validation, and then predict using the test sets.

- **Support vector machines** (SVM) converts features to points in high dimensional space and divide points from different categories by a gap as wide as possible for new feature classification.



- **Neural network** (e.g. MLP) consists of an input layer, an output layer and multiple hidden layers. It is a very power tool for text stance classification as it relies less on accuracy of feature extraction and can work on some crude features.



- **Softmax**: a generalization of the logistic regression, which predicts the label based on the maximum probability that the data may belong to a certain class.
- **Multinomial naive bayes** defines a generative process for the data set and assume that for all features, $p(x_1|y=c)$, $p(x_2|y=c)$, ..., $p(x_n|y=c)$ are independent given a specific category label $y=c$. Due to these properties, it is often used in text classification problems.

Results and Discussion

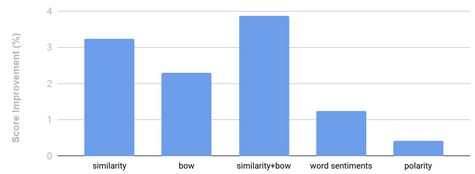


Figure 1: Performance improvements after adding the following features: similarity, bow(bag-of-word, and sentiment features.

A\P	Agree	Disagree	Discuss	Unrelated
Agree	**147(MLP),**	0	1528	228
	116(SF)	4	1446	337
	378(MNB)	39	1246	240
	97 (SVM)	0	1513	293
Disagree	34	**0**	444	219
	27	**0**	381	289
	62	**11**	436	188
	12	**0**	420	265
Discuss	198	0	**3761**	505
	122	0	**3556**	786
	625	38	**3168**	633
	86	0	**3691**	687
Unrelated	0	0	304	**18045**
	7	0	168	**18174**
	70	0	1096	**17183**
	6	0	202	**18141**

Table 1: Test set score by MLP, Softmax(SF), Multinomial Naive Bayes (MNB) and support vector machine (SVM)

Table 2: category assignment by running models on sub-categories

Label \ Predict	Related	Unrelated
Related	6460	604
Unrelated	614	17735

Label \ Predict	Agree	Disagree
Agree	1903	0
Disagree	697	0

Label \ Predict	A/D	Discuss
A/D	327	2273
Discuss	304	4160

- **Single Model:** < 20% accuracy on predictions of "agree" or "disagree" stances.
- **Running models on sub -categories** yields better predictions on "agree" and "disagree" stances

Therefore, two types of model combinations are considered.

- **2-model Combination**
  M1 - Classify related and unrelated stances
  M2 - For related stances, classify agree, disagree, and discuss
- **3-model Combination:**
  M1 - Classify related and unrelated stances
  M2 - For related stances, determine whether the stance is neural(discuss) or not
  M3 - For non-neural stances, determine whether the stance is agree or disagree.

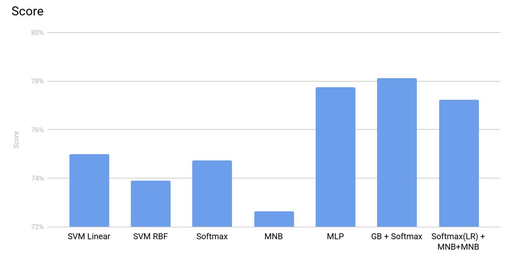


Figure 2: Comparison of test set score generated from different models. GB = Gradient Boosting Classifier. LR = Logistic Regression. MNB = Multinomial Naive Bayes. M_1+M_2(+M_3)= 2(3)-model combination.

Future work

1. Improve prediction of agree vs disagree.
2. Feature extraction based on domain knowledge

References

[1] Fake News Challenge. http://www.fakenewschallenge.org
[2] Sobhani P, etc. (2016) Detecting stance in tweets and analyzing its interaction with sentiment. *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics* 159-169
[3] Mohammad S.M. and Turney P.D.(2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34.