Detecting Emotional Scenes from Video Subtitles

Guide: Prof. Amitabha Mukherjee

March 31st, 2015

Group 6

Utsav Sinha

Rajat Kumar Panda

Problem Statement Background

Multimedia expresses emotional content using

- facial expression
- Dialogue
- the way of speaking
- the context
- background scene
- Music

An unsupervised model based on a mixture of these parameters can be used to automatically find emotional scenes of a video

Problem Statement

- To classify dialogues in a movie by tagging each dialogue with one of 5 emotions – happiness, anger, surprise, fear and disgust
- apply Natural Language Processing (NLP) techniques on subtitles of the video to achieve this goal

Word2Vec

- Word2vec provides an N dimensional vector for each word in its training corpus.
- The vectors are built using skip-gram model
- neural network implementation of Word2vec learns the context of words from sentences provided as untagged training data

Approach

- Word vectors would be obtained from training unlabeled Subtitle corpus (5000 videos) using Word2vec.
- Few subtitles (8-10) would have each dialogue hand labeled with one of the emotions. This acts as the ground truth.

Approach

To obtain the emotion of a dialogue a simple approach is to :

- Take the sum of all word vectors and finding the average vector
- Calculate the distance of this vector from the vector of 5 major emotions.
- The emotion of the dialogue is the one whose distance from the average vector is the minimum.
- If this minimum distance is more than a certain threshold, we can tag the dialog as emotionless.

Approach

- But the above model does not get any training from our labeled data. It just classifies without any learning
- So, we will use neural network (NN) to learn the function that maps word vectors (obtained from word2vec) to emotional labels.

Approach: SentiWordNet

- Another modification is to re-align the word vectors by incorporating extra dimensions of emotions to each word
- These extra dimensions can be obtained from synonym sets provided by SentiWordNet
- This process will help to bring together emotional words such as “pleasant”, “delight”, “cheerful” closer together to the major emotion of “happiness”.

Approach: SentiWordNet

- This step is useful since word2vec requires a huge corpus to train to bring out the context
- Also, word2vec is more generic than the goal of classification based on emotions alone. So vectors of similar emotion words may deviate far away.
- Most importantly, word2Vec keeps vectors close together based on context So nearest neighbors of word “happy” are:
- Unhappy, Terrible, Grateful, Pleased, Disappointed
- Clearly, Unhappy does not fit to be the closest neighbor of Happy in terms of Emotions

Approach

- The realigned vectors would then be similarly trained to find the mapping function using NN
- These 2 approaches, with and without SentiWordNet can then be compared for accuracy on a test data of few subtitle files

Addition

- Term frequency–inverse document frequency (tf–idf) can be used to remove stop words like “it”, “him”, “for” etc before NN is invoked
- This is useful since these stop words do not contribute to the overall emotion of a dialogue

Testing

- We hand labeled each dialogue of movie “Titanic” into one of happy, fear, anger, surprise, disgust, emotionless
- We then tested the simple approach of averaging word vectors to find the sentence vector
- This vector was classified into one of the 6 categories

Preliminary Results

Emotion	Ground Truth	Implementation	True Positive
Happy	385	34	31
Fear	310	121	50
Anger	112	227	35
Surprise	325	95	47
Disgust	157	659	82
Emotionless	757	910	528
	2046	2046	773

Accuracy = 773/2046 = 37.8%

Accuracy without emotionless dialogues = (773-528)/(2046-757) = 19.1%

Inference Drawn

- Since training was done on a small corpus, so word vector generated of less frequent words like “disgust”, “anger” were not accurate (vectors had smaller norms) as compared to more frequent words like “happy”, “good”
- So when calculating distance from average sentence vector, more dialogues had smaller norms and hence were classified as “disgust” or “emotionless”
- Results were poor since no learning on labeled data was done

How to Improve

- The training corpus should be increased in size.
- Even after that, words like “disgust”, “anger” would still have a relative frequency less than that of “happy”, “good” because of their usage in movie dialogues
- So tf-idf should be employed
- Stemming of words should be done

References

- Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christo-
- pher D Manning, Andrew Y Ng, and Christopher Potts.
- Recursive deep models for semantic compositionality over a sentiment tree-bank. In Proceedings of the conference on empirical methods in natural language processing (EMNLP) volume 1631, page 1642. Citeseer, 2013
- Seung-Bo Park, Eunsoon Yoo, Hyunsik Kim, and Geun-Sik Jo.
- Automatic emotion annotation of movie dialogue using wordnet.
- In Intelligent Information and Database Systems, pages 130-139.
- Springer, 2011