Word Embeddings

In our thesis we use Glove, Word2Vec as our main word embedding and along with some of modifications

1. Word2Vec

The model takes input as a large corpus of data and produces a vector space. [Word vectors](https://en.wikipedia.org/wiki/Word_vectors) are positioned in the vector space such that words that share common contexts in the corpus are located close to one another in the space.

Word2vec can utilize either of two model architectures to produce a [distributed representation](https://en.wikipedia.org/wiki/Distributed_representation) of words: [continuous bag-of-words](https://en.wikipedia.org/wiki/Continuous_bag-of-words) (CBOW) or continuous [skip-gram](https://en.wikipedia.org/wiki/Skip-gram). In the continuous bag-of-words architecture, the model predicts the current word from a window of surrounding context words. The order of context words does not influence prediction ([bag-of-words](https://en.wikipedia.org/wiki/Bag-of-words) assumption). In the continuous skip-gram architecture, the model uses the current word to predict the surrounding window of context words. The skip-gram architecture weighs nearby context words more heavily than more distant context words. CBOW is faster while skip-gram is slower but does a better job for infrequent words.

Word2Vec only take local contexts into account, does not take advantage of global count statistics.

1. Glove

The model is an [unsupervised learning](https://en.wikipedia.org/wiki/Unsupervised_learning) algorithm for obtaining vector representations for words. This is achieved by mapping words into a meaningful space where the distance between words is related to semantic similarity.

Training is performed on aggregated global word-word [co-occurrence](https://en.wikipedia.org/wiki/Co-occurrence) [statistics](https://en.wikipedia.org/wiki/Statistics) from a corpus.

Glove takes global information into account while learning dimensions of meaning and can be used to find relations between words.

1. Emotional Word Embedding Modification

According the original paper (Seyediatabari et al., 2019), the retrained models perform better than their original counterparts from 13% improvement for Word2Vec model, to 29% for GloVe vectors. Since we are using the word embedding for emotional classification purpose, we expect the word embedding to aid our deep learning model, improving the model’s overall performance.

Reference:

<https://en.wikipedia.org/wiki/Word2vec>

<https://en.wikipedia.org/wiki/GloVe_(machine_learning)>

<https://mlexplained.com/2018/04/29/paper-dissected-glove-global-vectors-for-word-representation-explained/>

Armin Seyeditabari, Narges Tabari, Shefie Gholizade, Wlodek Zadrozny. 2019. Emotional Embeddings: Refining Word Embeddings to Capture Emotional Content of Words. arXiv preprint arXiv: 1906.00112.