Sentiment analysis on Twitter

The data provided for this task is a corpus of tweets. As Twitter is an informal social media platform, the data has a lot of noise from elongated words to emoticons (emojis) to hash tags. Therefore, a lot of preprocessing was required to clean up the data so that it could be used for sentiment analysis.

Preprocessing

For the preprocessing part, I have used a lot of the regex code from the previous assignment. The first and foremost task was load up and preprocess the data. For preprocessing the following tasks have been performed:

1. Lower case all the data
2. Replace all URLs in the tweets with a word called “**urllink**”. The regex used for this is **'\b(http)(s)?:\/\/((([\w\/])(\.)?)+)\b'.** With hindsight, the replacement word should have been **<url>** as the latter is included in the Glove word embedding.
3. Replace user mentions with the “@” sign with a word **usermention**. The regex used for this is **^(?!.\*\bRT\b)(?:.+\s)?@\w+**. With hindsight, it would have been better to use **<user>** as a replacement word as the latter is contained in the Glove word embedding.
4. To remove short noisy words such as “a” and “I” which carry very little meaning in terms of sentiment, all one word words have been removed using regex, **\b[A-Za-z0-9]{1}\b**.
5. As lots of emojis are often used in social media, I have made some attempt to convert the happy and sad face emojis to **happyface** and **sadface** words. With hindsight, it would have been better to convert these directly to happy and sad words. The two regexes are given below:

**Happy: :\)|:]|:3|:>|8\)|\(:|=\)|=]|:\'\)**

**Sad: :\(|:\[|:<|8\(|\(:|=\(|=\[|:\'\(|:-\(**

Additionally, I have used the **WordNetLemmatizer** with default PoS tagging to lemmatize the words. This has helped reduce the vocabulary a little.

Feature extraction

* Lexicon with sentiment values
* NGrams
* Glove twitter word embeddings

Classifiers

* Lexicon rule based classifier
* Ngrams naïve bayes
* Word embeddings with SVMs

Performance