Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data

- 1911020, Kritarth Jain

Introduction:

- In the big data era, data is made in real-time thus businesses can utilize this ever growing volume of data for the data-driven decision-making.

- Social media data are unstructured hence the study proposes an effective text data preprocessing technique and develops an algorithm that weights the sentiment score in terms of weight of hashtag and cleaned text.

- The sentiments can be utilized to give companies insights into their products and also analysis of different competing products.

Sentiment Analysis:

- Sentiment Analysis is the process of categorizing opinions expressed in text, to determine whether the writer's attitude towards a topic or product is positive, negative, or neutral.

- Through Sentiment analysis companies can adjust to the present market situation and satisfy their customers in a better way.

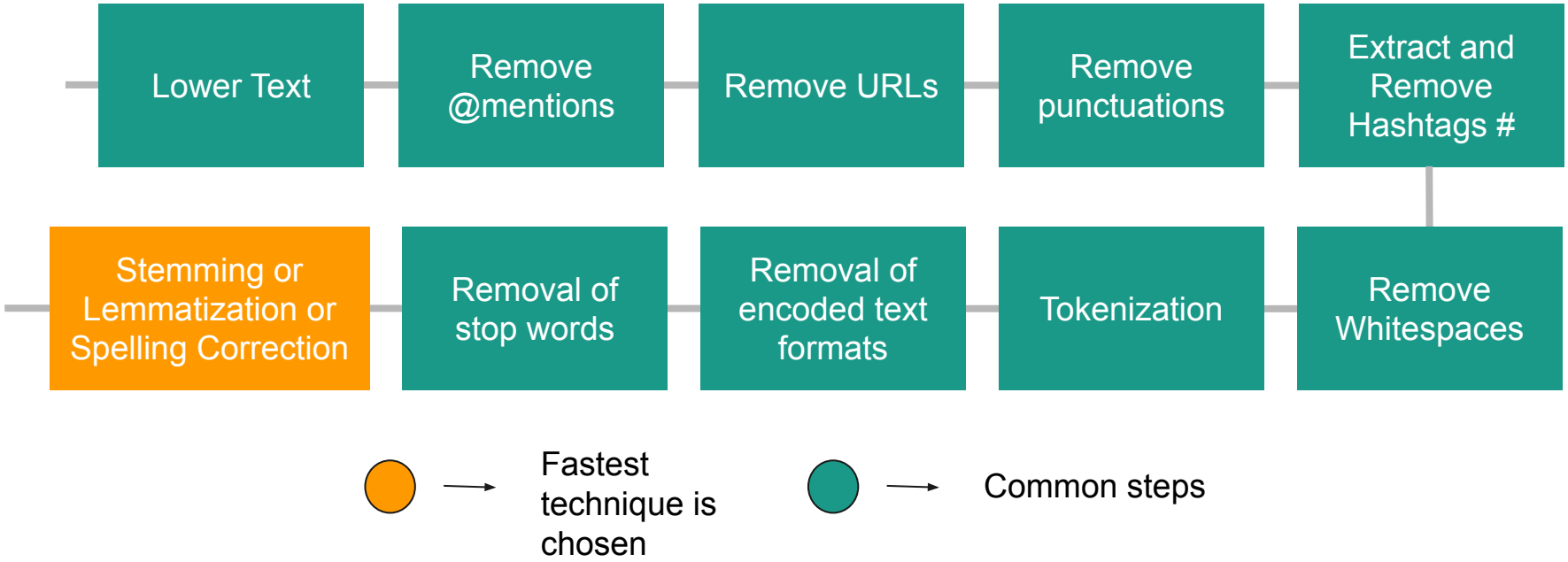
- Keeping the negative sentiments in mind companies can develop more appealing marketing strategies to improve brand status.

Importance of Preprocessing techniques:

- Preprocessing is used to produce the correct sentiment for effective decision making. It is implemented to remove the unstructured nature of data obtained from tweets.

- Many words in the text do not have an impact on the sentiment of the text. Keeping those words increases the dimensionality of the problem and makes the classification more difficult since each word in the text is treated as a dimension.(curse of dimensionality)

Preprocessing pipeline:



Comparison of Preprocessing techniques:

- **Stemming** : Stemming is the process of reducing a word to its word stem by removing suffixes or prefixes. The root word may or may not be a valid word. Eg: troubled,trouble,troubling are stemmed to troubl.

- **Lemmatization** : Lemmatization is very similar to stemming but the root word after lemmatization is a valid word and has similar meaning.
  Eg: troubled,trouble,troubling are lemmatized to trouble.

Comparison of Preprocessing techniques:

- **Spelling Correction** : The words are first passed for the spelling correction, where the spelling is checked and corrected if necessary. Then it is passed on to lemmatization.

- The text is first passed through the common process and then the faster algorithm out of out of stemming, lemmatization and spelling correction is used .Spelling correction takes much longer as compared to stemming and lemmatization, hence is not commonly used.

Proposed Algorithm:

- After preprocessing we would have the cleaned text and the hashtags if any which would be vectorized using TF-IDF vectorizer and would be given to the classifier for computing the sentiment.

- If there is no hashtag present then we compute the sentiment for the clean text only and report the result. But if hashtags are present we compute the sentiment as: **Tw = aH + bT** , **a** is weight for hashtags,**H** is score for hashtags, **b** is weight for clean text,**T** is score for cleaned text and **Tw** is final sentiment.

Proposed Algorithm:

- The weights **a** and **b** are hyperparameters which can be tuned for better results. The paper uses a = 0.4 and b = 0.6. If there are no hashtags present then a = 0 and b = 1.0 .
- **Classifiers Used**:
  1. Naive Bayes
  2. SVM
  3. Neural Networks

- According to the requirement, the classifier with most accuracy or the classifier with the least computation time can be chosen.

Conclusion:

- Businesses can use sentiment analysis for improving their products.
- Effective Preprocessing techniques are required to improve performance of classifiers for more accurate sentiment analysis.
- Hashtags also have meaningful information and hence must be used in addition to preprocessed text to produce better sentiments.

Strategy for implementation:

- Apple’s twitter sentiment dataset would be used for training the classifiers.
- Preprocessing of tweets would be done using regex and nltk libraries in python.
- Sklearn would be used for making naive bayes and SVM models.
- GetOldTweets3 package would be used to get tweets for sentiment analysis on Google Now and Amazon Alexa.

References:

- Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data
- The Role of Text Pre-processing in Sentiment Analysis
- https://towardsdatascience.com/applications-of-sentiment-analysis-in-business-b7e660e3de69
- https://www.datacamp.com/community/tutorials/stemming-lemmatization-python