Writing Scientific Papers and Software

Beril Besbinar, Dimitrios Sarigiannis, Panayiotis Smeros
Department of Computer Science, EPFL, Switzerland

***Abstract*—**

I. Introduction

II. Preprocessing Techniques

In this section we describe the preprocessing techniques that we used in order to homogenize the raw tweets and emphasize words and writing styles that denote sentiment. After applying these techniques we were able to emphasize writing styles and words that denote sentiment (e.g., emojis, repeating the last character of a word, etc.) as well as handling most of the spelling errors and slang words usage (e.g., with lemmatization and stemming).

The methods that we implemented were inspired from the respective methods of *GloVe* [1] and are summarized as follows:

A. Contractions Expansion

In informal speech, which is widely used in social media, it is common to use contractions of words (e.g., don't instead of not). This may result in misinterpreting the meaning of a phrase especially in the case of negations. For this reason we implemented a method that expands most of these contractions.

B. Emojis Transformation

Emojis are widely used in social media and they can alone give a good estimate of a post's sentiment. In order to handle the different writing styles of emojis (e.g., :), :-) and (: are all smile faces) we transform all of them into the word that describes better the sentiment that they denote. Thus we have smile, lol, neutral and sad faces as well as hearts.

C. Emphasize Repeated Punctuation

Some punctuation marks like exclamation point (!) denote sentiment especially when they are repeated within a tweet. With this method we emphasize that repetition.

D. Emphasize Repeated Last Characters

Similarly to the punctuation marks, the repetition of the last character of a word reveals sentiment (e.g., *I am happyyyy*). We introduce a method that corrects the spelling of such words (e.g., converts *happyyyy* to *happy*) and adds a special tag to indicate the repetition.

E. Filter Numerical Expressions

Since numbers in general do not hold any special sentiment semantics, we replace all the numerical expressions with tags that stand for the existence of such expressions.

F. Split Hashtags

Hashtags are one or more words concatenated with the symbol #. Some of the tokens of the hashtags may hide useful information for the sentiment of a post (e.g., #lovemyjob). Splitting a hashtag into token is a difficult problem [2] since we may have multiple, ambiguous splits (e.g., #homestore can be split into either *home store* or *homes tore*). Having typos and using slang words makes it even more difficult. In order to solve this problem we used a dictionary with the most frequently used English words in descending order (according to Zipf's law) and tried to guess the correct split.

G. Emphasize Sentiment Words

Similarly to emojis, some English words denote clearly an emotion (e.g., *anxiety* or *happiness*). In order to emphasize the existence of such words we were advised by a lexicon with positive and negative words [3].

H. Part-Of-Speech Tagging

Part-Of-Speech tagging helps us understanding the structure of an input text in order to discover its most important words. Since posts in social media are informal, without following many grammatical rules, in most of the cases this method did not help.

I. Lemmatization and Stemming

In order to homogenize verbs being in different tenses we applied lemmatization and stemming techniques to the words of the input tweets.

J. Stop-words Filtering

Stop-words are very common English words that can be found in almost every tweet and thus do not imply any sentiment (e.g., the articles *The*, *A* etc.).

III. Representations of Tweets

IV. Model Selection

V. Conclusions

References

[1] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[2] S. Khaitan, A. Das, S. Gain, and A. Sampath, "Data-driven compound splitting method for english compounds in domain names," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM '09. New York, NY, USA: ACM, 2009, pp. 207–214. [Online]. Available: http://doi.acm.org/10.1145/1645953.1645982

[3] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.