United We Stand: Tweet Pooling for a better LDA

*Author*1[1] *Author*2[2]
(1) INSTITUTE_1, address 1
(2) INSTITUTE_2, address 2
mail-id, mail-id

Abstract
DEMO - Social networks such as Facebook, LinkedIn, and Twitter have been a crucial source of information for a wide spectrum of users. In Twitter, popular information that is deemed important by the community propagates through the network. Studying the characteristics of content in the messages becomes important for a number of tasks, such as breaking news detection, personalized message recommendation, friends recommendation, sentiment analysis and others. While many researchers wish to use standard text mining tools to understand messages on Twitter, the restricted length of those messages prevents them from being employed to their full potential.

Keywords: Topic Modelling, LDA, Tweets.
Keywords in L_2:
.

---

1 Introduction

Some points to be covered here:

- Twitter introduction and problems faced while analysing tweets
- Motivation for Topic Modelling for tweets: information needs
- Difficulties faced by general LDA on tweets
- How we address these issues and our contribution

2 Topic Models : Latent Dirichlet Allocation

A brief description of LDA.

3 Pooling Schemes

In this section we describe various tweet pooling schemes in detail.

1. Burst Score wise
   - Definition of Burst Score
   - Pooling by Burst Score
2. Author-wise Pooling
3. Temporal Pooling

    - 30 minutes
    - Hourly
4. Conversational
5. Hashtag-wise

4 Lexical Normalization of Hashtags

In this section we describe hashtag normalization in the following 3 subsections.

4.1 Scoping Hashtag Normalization

- Task definition
- Figures / analysis of the types of hashtags we encounter in a dataset.

4.2 Candidate Set Generation

- Letters
- Numbers
- Letters+Numbers
- Abbreviations
- Common prefix/suffix - substring

4.3 Context Based Pruning

For each of the hashtags obtained from the above method we compare the tweet collection of each hashtag with the hashtag in question (using the number of common words as a basic metric) and based on some filtering score we select the top candidate.

5 Experimental Setup

5.1 Dataset Description

Here we describe the datasets we create in detail:

- Generic
- Specific
- WDS Conference
- other

5.2 Document Characteristics after Pooling

A big table which describes document characteristics for each pooling scheme in each dataset.

5.3 Evaluation Metrics

We test the different pooling schemes on a variety of metrics for different goals.

1. Purity
    - to see which schemes is best suited to reproduce known clusters
2. NMI
3. PMI
    - to measure which scheme gives the most coherent topics
    - We modify this metric so that instead of external source it uses the tweet dataset itself to calculate the probabilities.

6 Results

1. Purity
    - a single bar graph with Pooling scheme on X-Axis and Purity Score on Y-axis for all the 3 datasets
2. NMI
    - a scatter graph with #topics on x-axis and NMI score on y-axis
3. PMI
    - a single bar graph with Pooling scheme on X-Axis and PMI Score on Y-axis for all the 3 datasets

7 Observations

Herewe list down pointwise our observations, the affect of various settings and our main findings/insights.

8 Related Work

Here we discuss in detail the prior work done on Topic Modelling for tweets, majorly dividing it into 4 categories:

1. Topic Models for tweets
2. Visualizing Topic Models
3. Evaluating Topic Models
4. Application of Topic Models for twitter tasks(events, earthquake, et cetera)

9 Conclusion

Acknowledgments