Automatic News Generation Based on Twitter

Ali Alavi[1], Rolf Jagerman[1], and Tsay Kai-En[1]

ETH Zürich, Zürich, Switzerland
alavis@ethz.ch, {rolfj, tsayk}@student.ethz.ch

1 Abstract

In this report, we describe our approach towards generating news topics based on twitter data. We tackle this problem by running a stochastic gradient descent classifier on a large set of news articles collected from different news agencies, and then using this classifier to classify twitter posts into three news categories: sports, politics, and technology. Then we tokenize these classified tweets in order to extract names and nouns used in each tweet. Finally we perform a time series analysis on these set of words and recognize top ten trending topics as newsworthy. We present the results in a web based interface. The results, especially in politics and sports category, bear resemblance with the trending topics as reported by news agencies.

2 Introduction

The authors were motivated by a simple question: can we generate more accurate and less biased news using twitter data in comparison to traditional news agencies? Would such a system have a potential of becoming an alternative to mainstream news sources? If so, such system can be a more trustable source of unbiased news, which in turn will have immense effect on public awareness, knowledge and discourse.

Although there are many tools and websites, such as Google News, which automatically aggregate and present news articles, their data sources are mainstream news agencies. We, on the other hand, want to use public posts as our data source, hence using collective knowledge of citizens as our news agency. In this model, every twitter user can play a small, yet collectively significant role in news gathering, and hence in generating valuable news articles, even without his or her knowledge. Effectively, the audience will become the content provider as well.

3 Performance Measurements and Results

4 Conclusion