# Making the news concise again

Saurabh Verma
Indian Institute of Technology Roorkee
saurrv@gmail.com

Vikash Kumar
Indian Institute of Technology Roorkee
vikash4466kumar@gmail.com

Sachin Aggarwal
Indian Institute of Technology Roorkee
sachinaggarwal077@gmail.com

Balasubramanian Raman
Indian Institute of Technology Roorkee
balarfcs@iitr.ac.in

***Abstract*—Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.**

***Keywords*—Deep learning for text, news, redundancy, content popularity.**

## I. Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. [1]

## II. Business

- **Content Caching and Traffic Management** There is a hidden cost to publishing content, the cost to review and maintain the content. The millions of articles also affect the usability and maintainability of the site. In the long run, it is necessary to tackle redundant, outdated and trivial content which has been cursing the site.
- **Advertising** This work finds it's use in media advertising and ad placement.
- **News Aggregation**
- **Trends Forecasting**

## III. Related Work

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

## IV. Solution Framework

### A. Preprocessing

### B. Text Understanding

### C. Clustering

### D. Popularity Prediction

## V. Experimental Evaluation

### A. Dataset

1) **AG's news corpus** We obtained AG's corpus of news articles on the web[1]. It contains more than 1 million news aricles from more than 2000 sources. We use the source, title, description, rank and pubdate fields for our experiments.
2) **Financial News Dataset** We obtained financial news dataset of Bloomberg and Reuters[2]. It contains 450,341 articles from Bloomberg and 109,110 articles from Reuters. It provides us title, content, date and author of the article.

### B. Baseline

## VI. Conclusion

In this paper, we improve the quality of news cache and recommendations by predicting popularity of articles prior to publishing. The need for the same arises from the stiff competition among different news agencies and aggregators. Through deep convolutional neural networks, we extract features of articles at character-level. To remove redundant information, we make highly specific clusters of news items. Finally, we predict the most popular pieces in diffrent clusters to provide the set of most popular articles, which is then used for multiple use-cases in content caching, advertising, forecasting and

[1] https://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
[2] https://github.com/philipperemy/financial-news-dataset

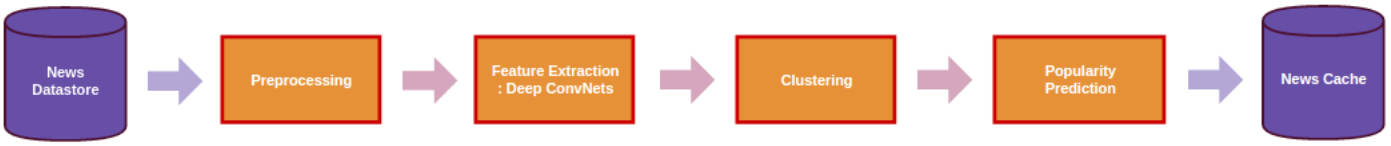


Fig. 1: Solution Framework

recommendation. With an initial survey, we ensure inceptive results of the pipeline versus different competitors. Lastly, we compare with different baselines to ascertain quality of our work.

## VII. Future Work

1) Information explosion is prevalent not only in news items, but across different content classes. There rises the opportunity to extend the current work for multiple classes.
2) There are abundant works [2] [3] which when given initial popularity of some content, are better able to forecast popularity. With that notion, we want to adapt the predictions to incoming hits.
3) With growing amount of personalization in different news agents [4] [5], there exists a demand and an opportunity for the same in this work.

## References

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[3] A. Tatar, J. Leguay, P. Antoniadis, A. Limbourg, M. D. de Amorim, and S. Fdida, "Predicting the popularity of online articles based on user comments," in *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, (New York, NY, USA), pp. 67:1–67:8, ACM, 2011.

[4] I. Ilievski and S. Roy, "Personalized news recommendation based on implicit feedback," in *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge*, NRS '13, (New York, NY, USA), pp. 10–15, ACM, 2013.

[5] J. Bao and M. F. Mokbel, "Georank: an efficient location-aware news feed ranking system," in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 184–193, ACM, 2013.