Thesis Proposal : Information/News exposure & issue salience

Written by: Ziyad Meftah
Supervised by: JProf. Dr. Claudia Wagner & Dr. Juhi Kulshrestha

Introduction:

As we all know, the technology and the way we communicate have evolved tremendously in recent decades. We have a whole range of digital devices that allow us to consult the media and access a wealth of information. However, the way we see and consume this information differs from an individual to another based on different aspects (age, gender, occupation, location and most importantly preferences).

In this project, we focus in particular on news articles and how people consume them with respect to their preferences. Do people really stick to their preferences and read articles that are closely related to the socio-political issues they claim to be most important to them ?

This leads us to the main research question which is : "***How is the issue that a person considers most important related to what they browse?***"

At the end of this project, we will provide a detailed explanation of the method used for identifying topics. In addition, we will share usable notebooks with our scripts and our approach on creating browsing profiles.
We might go further and analyse the findings in correlation with other aspects such as age or gender.

Materials and methodology:

To conduct the study, we use web tracking data from which news articles browsed by the participants were scraped, parsed and ready for analysis. In addition, we use results of two surveys conducted on the participants on two different occasions (April 2019, May 2019) in which they specify the issue they consider to be most important from a set of predefined issues (Annex 1).

Now, the task will be identifying the main topic of each read article and map it to the predefined issues. That way, we can create profiles from the clustered topics.

The main problem here was choosing the most adequate methodology. After reading multiple papers, we shortlisted papers that are most relevant to our idea (Relevant papers), we found out there are two possible ways for applying text classification on our data. There is the ***supervised learning*** technique that requires a dataset of pre-labelled articles to use as training data and performing a feature selection task to reduce the dimensionality of our classification model using either chi square or Information gain. The model itself can vary depending on which setting we're in. For the single label setting, the decision trees or random forest techniques are widely used. As for the multi-labeling setting, Binary Relevance or Classifier Chains are used, which are nothing but an ensemble of trained single-label binary classifiers, one for each class. Each classifier predicts either the presence or the non-presence of one class. The union of all classes that were predicted is taken as the multi-label output.

The limitation of the supervised learning lies in that using a static dataset prevents from capturing new keywords for certain article topics which will result in bad classification and also the creation of the dataset will require a lot of hand-annotating of the articles with the survey-categories.

The ***unsupervised learning*** technique is the one we opt for. We will apply a *topic modeling approach*. The approach is to use the **Latent Dirichlet Allocation** model which basically clusters documents based on word usage in order to get the main topics that occur in a set of documents (articles). This approach has multiple variants and we would like to evaluate the performance of some of them. We are interested in **LDA2vec,** a technique that aims at mixing the best of LDA and **word embeddings** which is a word representation that allows words with similar meaning to have a similar representation. This provides a solution to the problem of dimensionality linked to the size of dictionaries. We are also interested in the ***semi-supervised learning*** **guidedLDA** which can incorporate the predefined issues as priors to force the LDA model to give topics that are similar to the ones we already have. Its limitation is that we need to create sets of seed words, meaning words that are most representative of the underlying topics in our articles which is a hard thing to do since we don't know how the topics are represented in our corpus. We can try keywords related to the issues instead. These will be extracted manually from wikipedia or news articles themselves.

The expected outcome will be exclusive clusters of articles. We can then take two approaches. First, we could manually label each cluster to a meaningful topic or issue(s). Or second, we could use the clusters as the reduced latent dimensions. So we would create topic profiles/vectors for each article. Then aggregate these profiles to a user to capture the articles they browse. Finally, we link these user topic profiles with the issues that they consider most important to compare across issues.

Related work:

A lot of work has already been done in identifying predefined socio-political issues in news articles. The most occuring idea was to use boolean search which is nothing but looking whether a word is present or absent in a document. For that, W. R. Neuman, L. Guggenheim, S. M. Jang, and S. Y. Bae 2014 created, for each issue, a set of key identifying terms or phrases unique to that issue but didn't specify how these words were identified. In the political setting, X. Yang, B.-C. Chen, M. Maity, and E. Ferrara used the same technique by manually looking for the most frequent words that could be indicative of specific topics and sound meaningful to ordinary readers. Similar to that, Y. Kim, W. J. Gonzenbach, C. J. Vargo, and Y. Kim built their issue's keywords based on lexicon-based lists from previous agenda-setting research. Another approach was introduced by H. Kwak, J. An, J. Salminen, S.-G. Jung, and J. Jansen 2018 where they built a set of 11 common topics and collected co-mentions using word embeddings (i.e., topics that are mentioned together with any of the daily top 100 popular topics per country).Then, for each co-mention, they computed the distance from each topic and assigned the topic to the shortest distance. Finally, and in a semi-supervised setting, Yamshchikov, Ivan P., and Sharwin Rezagholi 2018 trained a set of (7) binary text classifiers based on convolutional neural networks on annotated sentences and applied these classifiers to the non-annotated sentences. Also, A. Bilbao-Jayo and A. Almeida 2018 did the same thing by using convolutional neural networks with Word2Vec word embeddings for discourse classification where sentences are taken as inputs. Their model was trained using election manifestos annotated manually by the Regional Manifestos project.

Taking into consideration all the above, what makes our methodology stand out from the previous works is that it focuses on full-length articles as opposed to most of the previously mentioned papers that focus only on sentences such as tweets. Also, it doesn't involve training our model on annotated documents and can be used at any time without having to update the keywords of any topic.

References:

"W. R. Neuman, L. Guggenheim, S. M. Jang, and S. Y. Bae, “The Dynamics of Public Attention: Agenda-Setting Theory Meets Big Data,” J. Commun., vol. 64, no. 2, pp. 193–214, Apr. 2014."

"X. Yang, B.-C. Chen, M. Maity, and E. Ferrara, “Social Politics: Agenda Setting and Political Communication on Social Media,” in Social Informatics, 2016, pp. 330–344."

"Y. Kim, W. J. Gonzenbach, C. J. Vargo, and Y. Kim, “First and Second Levels of Intermedia Agenda Setting: Political Advertising, Newspapers, and Twitter during the 2012 U.S. Presidential Election,” Int. J. Commun., vol. 10, no. 0, p. 20, Sep. 2016."

"H. Kwak, J. An, J. Salminen, S.-G. Jung, and J. Jansen, “What We Read, What We Search: Media Attention and Public Attention Among 193 Countries,” Feb. 2018."

"A. Bilbao-Jayo and A. Almeida, “Political discourse classification in social networks using context sensitive convolutional neural networks,” in Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, Melbourne, Australia, 2018, pp. 76–85."

"A. Fang, P. Habel, I. Ounis, and C. MacDonald, “Votes on Twitter: Assessing Candidate Preferences and Topics of Discussion During the 2016 U.S. Presidential Election,” SAGE Open, vol. 9, no. 1, p. 215824401879165, Jan. 2019."

“Yamshchikov, Ivan P., and Sharwin Rezagholi. "Elephants, Donkeys, and Colonel Blotto." arXiv preprint arXiv:1805.12083 (2018).”

Annex 1:
Predefined set of issues

1 Crime **(UK)(US)**
2 Economic situation **(UK)(US)**
3 Rising prices / inflation / cost of living **(UK)(US)**
4 Taxation **(UK)(US)**
5 Unemployment **(UK)(US)**
6 Terrorism **(UK)(US)**
7 Housing **(UK)(US)**
8 Government debt **(UK)(US)**
9 Immigration **(UK)(US)**
10 Health and social security **(UK)(US)**
11 The education system **(UK)(US)**
12 Pensions **(UK)(US)**
13 The environment, climate and energy issues **(UK)(US)**
14 The decision of the United Kingdom to leave the European Union (Brexit) **(UK)**