Le Pham Minh Duc

Un-polarizing news in social media platform

Master’s thesis of mathematical information technology

September 30, 2018

Author:Le Pham Minh Duc (Lê Phạm Minh Đức)

Contact information: [miduleph@student.jyu.fi](mailto:miduleph@student.jyu.fi), minhduc1993@yahoo.com

Supervisors: Oleksiy Khriyenko (oleksiy.o.khriyenko@jyu.fi)

Title: Un-polarizing news in social media platform

Työn nimi:

Project: Master’s thesis

Study line: Web Intelligence and Service Engineering

Page count: x+y (x = page count without appendices; y = page count of appendices)

Abstract: Abstract…

Keywords: Keywords…

Suomenkielinen tiivistelmä: Abstract in Finnish…

Avainsanat: Keywords in Finnish…

Preface

This is where you can write a preface for your thesis. The preface is optional so you don’t have to write it if you don’t want. Also, the glossary on the next page is optional. List of figures and tables are required if there is at least one figure or table in the thesis.

**The most important guideline** for the use of this template: look from the title list which style has been used to which section of the document and use similar styles also in the future when modifying the document. Use always styles, never do formatting outside styles.

If you notice bugs or mistakes in the template, please inform them to Antti Ekonoja ([antti.j.ekonoja@jyu.fi](mailto:antti.j.ekonoja@jyu.fi)).

Jyväskylä, September 30, 2018

Le Pham Minh Duc (Lê Phạm Minh Đức)

Glossary

NLP Natural language processing

DCOM Distributed Component Object Model  
More explanation…

C++ Shouldn’t need any explanation…

List of Figures

Figure 1. Service architecture 4

Figure 2. Test figure 6

List of Tables

Table 1. Example of the table 4

Contents

[1 INTRODUCTION 1](#_Toc522644710)

[1.1 Problems overview 1](#_Toc522644711)

[1.2 Proposed solution and research questions 1](#_Toc522644712)

[1.3 Research method 2](#_Toc522644713)

[1.4 Thesis structure 2](#_Toc522644714)

[2 UN-POLARIZING ALGORITHM 3](#_Toc522644715)

[2.1 Natural language processing 3](#_Toc522644716)

[2.1.1 What is natural language processing 3](#_Toc522644717)

[2.1.2 Evaluation of available NLP services 3](#_Toc522644718)

[2.1.3 Stanford CoreNLP 3](#_Toc522644719)

[2.2 Article processing 3](#_Toc522644720)

[2.2.1 Data gathering technique 3](#_Toc522644721)

[2.2.2 Web content parser 4](#_Toc522644722)

[2.2.3 Article annotation 4](#_Toc522644723)

[2.3 Un-polarizing algorithm 4](#_Toc522644724)

[2.3.1 Sentiment analysis 4](#_Toc522644725)

[2.3.2 Named entity recognizer 4](#_Toc522644726)

[2.3.3 Article similarity calculation 4](#_Toc522644727)

[2.3.4 Un-polarizing algorithm 4](#_Toc522644728)

[2.4 Solution evaluation and related works 4](#_Toc522644729)

[3 SERVICE DEPLOYMENT 5](#_Toc522644730)

[3.1 Backend service deployment 5](#_Toc522644731)

[3.2 User interface and user experience design 5](#_Toc522644732)

[4 CONCLUSION 6](#_Toc522644733)

[Bibliography 7](#_Toc522644734)

[Appendices 8](#_Toc522644735)

[A Title of the first appendix 8](#_Toc522644736)

[B Title of the second appendix 8](#_Toc522644737)

# INTRODUCTION

## Problems overview

Ideological polarization has been a problem in our society for quite a long time. (that reference from 1986). With the rise of social media (citation here or not), it’s estimated that 66% of US citizen use social media as one source of news (citation from that web page), the amplification of ideological polarization has been increasing much faster than previously due to social endorsement, and other social media techniques that is used to keeps its user engaged (Sporh. 2017). This creates the echo-chamber effects that, by the design of social networks that only show what the users want to see, make the user even furthermore polarized into his own belief and makes him see the world wrongly, which may turn the user into some extremists that might be harmful for the society.

Scholars have been researching about this problem and solutions are proposed (many citations needed) but these solutions are either too impractical (e.g: needing the giant media companies to change their entire business models) or just way out of reach of the scholar scope (e.g: needing of the government’s intervention on the issue or people to stop using the services).

## Proposed solution and research questions

The main goal of this thesis is to find the way to break the people’s echo chamber that is mostly caused by the effect of social medias only show the user what he/she wants to see. To combat this, we need to show the user the news from the other side of view. If he/she reads about the opening of a new coal mines help creating a few hundreds of new jobs for the area, he should also know that the new coal mines will cause a great damage to the environment and might cause some local wild-life to disappear.

On top of that, the service must be accessible and easy to use, as the reason of many people using social media as their main source of news as it’s so convenience to have one place to go to and can see both your friend’s status as well as news.

With that goal in mind, the main research question of the thesis is:

* **How to find articles with alternative (different) points of view to a given article?**

We will only attempt to find the news that is relevant to the article but also provide oversight from different point of view that the first article misses. We will not check if the news is credible (but we will try to only provide news from credible sources) or if it is true, we simply provide the user different articles from many points of views so that he/she can choose to interpret it whatever he/she wants to.

Afterwards, with the first question answered, we will address two additional support questions on deploying the news un-polarizing service for the mass to use:

* **What is the most convenience way to deliver the service for the user?**

If the service is too complicated to use, or requires too many unnecessary steps, the user will rarely use the service at all, which will defeat the whole purpose of the system.

* **How to engineer the service so that it is autonomous, up-to-date and scalable?**

As a news service, it must always catch up with the latest news to be relevant to use.

## Research method

Abc, test text. Please fill in later.

## Thesis structure

This thesis will be divided into four main parts, including this introduction. The second chapter, named “Un-polarizing algorithm” will attempt to answer the main research question on how to find articles with alternative points of view to a given article.

As this chapter attempts to answer the main thesis question, including the theoretical approach as well as solution implementation from the ground, it thus can be furthermore divided into 4 smaller sub-chapters in which, the first sub-chapter starts with the fundamental technologies used for the works, the author original hypothesis and the reasons for choosing the services used in the thesis implementation part. The second sub-chapter focus on processing and storing the articles/news/documents retrieved from the internet using the technologies presented previously. With the articles processed and stored, the third sub-chapter provides the un-polarizing algorithm that answer the main thesis questions. Finally, the second chapter ends with evaluation of the un-polarizing algorithm as well as comparison to other existing solution.

The third chapter tackles the two subsequence researches question of providing the un-polarizing algorithm in the most convenience way for the general mass to use and how to develop the algorithm to be a cloud service that can support a large amount data if needed.

Finally, conclusions for the thesis as well as possible future work and extension are given in the last chapter.

# UN-POLARIZING ALGORITHM

## Natural language processing

### Original hypothesis

Our original hypothesis is that: given two articles talking about one similar topic (*for example: The US’s involvement in Iraq*), if one article has a positive views on the situation (*ie: Saddam Hussein is a terrible man and the people living under his reign are suffering badly*) and the other has a negative views regarding the same situation (*ie: it furthermore destabilizes the region and the main intention of the war was because of oil*). With articles telling about the same story but with different sentiment value, it could be interesting for the reader to see from different kind of attitudes about the same topic, thus, bring him to different point of views about a problem.

This hypothesis, however, after some implementation and evaluation, was proven to be not good enough (**more in-depth in chapter 2.3.2 Sentiment analysis**). The final solution of this thesis utilizes more complexed calculations and processing techniques that was not originally planned from the start. However, the initial hypothesis did create a solid technology base to start working on: Named entity recognizer (*to understand the article topic*) and sentiment analysis (*to understand the positivity/negativity of the article*).

### Natural language processing and evaluation of available NLP services

Named entity recognizer and sentiment analysis are two of the many sub-tasks covered by Natural language processing (NLP) technology, *which is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, how to program computers to process and analyze large amounts of natural language data. (Source from wiki- more information needed).*

For just named entity recognizer only and sentiment analysis only, there exists multiple tools that do not require NLP. For example, a simple named entity recognizer (*citation needed)* consists of a dictionary of many individual nouns and a lookup function to match the result from the document and the dictionary. Similarly, a basic sentiment analysis can be made by assigning a sentiment value to each of the adjectives in the sentence/paragraph and then calculate the sum of these values as the sentiment value for the sentence/paragraph. However, these methods are quite barebone and quite usually not really correct since their rules are quite flawed and there are many ways for them to misinterpret the true sentiment of the sentence. Sophisticated NLP tools with “sentiment analysis” supported have a more complex ways of defining the sentiment value

Named entity recognizer and Sentiment analysis can be easily created without NLP using a rule-based system. However, these basic tools are not good and sophisticated NLP tools are much better because they have a more complex system that take many other things into consideration.

### Stanford CoreNLP

Many big companies offered NLP services like Microsoft, IBM, Google, each with their own technology and strong/weakness. These tools, however are 1. Expensive to run, 2. Close source and are subjective to change in any moment (like IBM) which make them not attractive to use.

Stanford CoreNLP is an NLP tool created by Stanford university, free and open source and quite easy to setup/use and very powerful as it supports many features.

After some testing some of these tools, we decided to go with Stanford CoreNLP because Microsoft is quite weak and in-accurate,

IBM is expensive and is a pain to setup and they change the name/domain quite often so not reliable, even though their named entity recognizer and sentiment analysis are really good.

Google is ok, looking and feeling very similar to Stanford CoreNLP (like they took the source code of NLP and improved it upon). They strip a lot of features from CoreNLP but the things they keep (like NER and Sentiment) are much better than the stock version in CoreNLP.

Still, I would like to have everything in one place, free and will guarantee to work no matter what, so, I went with Stanford CoreNLP. The important part of this thesis is the algorithm behind the news unpolarizer, not from how can I choose a better service to use.

## Article processing

### Data gathering technique

Local

database

Web

content

processor

Article

annotator

Requested URL

News gatherer

News

suggestions

module

Suggested

articles

Figure 1. Service architecture

This thesis is fortunate in a way that all the data for this is just wildly available everywhere, not like other thesis where people have to search hard to find the correct data and to test if it fits their hypothesis or whatever.

For the data collection phase, when I first started, I just went to some news site everyday and copy some tops links I see from that. After few minutes, I would have 50 or so of links to add to a local text file, so the data collection is easy for me.

Late, for a better approach, there are a lot of news APIs available, for example, Google News or many other things, just one simple APIs and it can give you all the thing you want. It’s also good for evaluation later on, because these APIs allow you to search for query by word, date and time, which will be useful to compare the results between our stuffs and theirs.

Web content parser. The article on the internet are presented inside a web page, with just not only the article itself, but with tons of other unnecessary things like html tag for formatting, images and captions, links to other things on their website and advertisement text.

To strip away all the unnecessary function, I have a web content parser module which utilize a similar technique to reader mode on Firefox or Safari which will automatically strip away all the non-article part in the web content. However, this is not enough since it only strips the advertisement and related news, the core article and the html tag around it still persists, which, to solve, I wrote a smaller module to automatically remove all the html tag as well as the image and the caption, which is not really accurate since each website have different layouts and ways to present content, make parsing out the content really not effective, since, for example, some website, when they end the sentences or the caption of the image, the don’t add the “.” (dot) or sometimes they make multiple dots, which make finding out the article contents with proper sentences are quite a problem in most of the article, which make the overall annotating result worse.

One effective way of ensuring that the sentences forwarded into the CoreNLP annotator are correct is to use a tools call SMMRY, an article summarization tools, which will read through the article and gives out the sentences that it thinks contains the most important information of the article. This tool is quite effective for our case as it first, strip away all of the garbage contents like html, tag and about this website … which make the article annotation work correctly. One downside of this is that it will not return the whole article, only part of it (the more important part) so there might be possible information that will be lost during the stripping of the content. Still, I think it’s better to use this SMMRY tool instead of my own web content parser because it’s not very good and accurate.

When all the data-collecting and parsing is completed, we push the parsed data into our article annotator.

### Article annotation

Article annotation!

## Un-polarizing algorithm

### Sentiment analysis

### Named entity recognizer

### Article similarity calculation

|  |  |  |  |
| --- | --- | --- | --- |
| Word | Year | Magnitude | Example |
| example | [1700,2000] | [1,10] | example |
| example | [1950,2000] | [1,106] | example |
| example | [1995,2000] | [10–6,106] | example 1, example 2 |

Table 1. Example of the table

### Un-polarizing algorithm

## Solution evaluation and related works

# SERVICE DEPLOYMENT

## Backend service deployment

## User interface and user experience design

Local

database

Web

content

processor

Article

annotator

Requested URL

News gatherer

News

suggestions

module

Suggested

articles

Figure 2. Test figure

# CONCLUSION

Hope you enjoyed the text...

In the bibliography the recommendable style is Chicago. You can also use other styles: the main thing is that the styles of the bibliography and referring technique are **consistent** in the whole thesis.

Bibliography

Spohr, Dominic. "Fake News and Ideological Polarization: Filter Bubbles and Selective Exposure On Social Media." *Business Information Review* 34.3 (2017): 150-160.

Meikäläinen, Matti. *Tietotekniikan perusteet.* Helsinki: WSOY, 2000.

Santanen, Jukka-Pekka. "Opinnäytteiden kirjoittaminen, lyhyt oppimäärä." 2000. http://users.jyu.fi/~santanen/info/kirjoittamisesta.html (accessed 5.10.2012).

Writer, John. "Name of the article." *Computer*, no. 2/2000 (2000): 31-32.

Appendices

1. Title of the first appendix
2. Title of the second appendix