FIS | VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE
FAKULTA INFORMATIKY A STATISTIKY

fis.vse.cz

ENGAGEMENT OF CZECH MEDIA ON FACEBOOK

Antonín Pavlíček

Department of System Analysis

Faculty of Informatics and Statistics

University of Economics Prague

Introduction

- Traditional media share their posts and articles on social media to stay connected with their audiences
- Since television and print are becoming less popular, especially among the young generation, Facebook is often becoming the only source of information and news they have.
- The main goal of this research is to define the online reader base of 8 chosen Czech media based on the comments under article posts on Facebook.
- We will also answer 12 hypotheses created based on common generational and gender stereotypes and prejudices from everyday life and the current pandemic.

Methodology

- This paper uses the Reuters Digital News Report 2018 to select 16 most familiar Czech online media => (Seznam.cz, iDnes.cz, Aktualne.cz, Novinky.cz, Czech Television news online, TN.cz, iPrima.cz, Bleska.cz, Denik.cz Super.cz, iHned.cz, Lidovky.cz, Reflex.cz, DVTV.cz, Tyden.cz, and iRozhlas.cz)

Web Scraping

- We used the web scrapping method to extract data from Facebook
- we collected around 170,000 data from Facebook users commenting on selected Czech media groups about first name, last name, sex, and age divided into categories, region, and Facebook fan batch.
- We downloaded around 2000 records for each media for each of six media pages which makes 16693 records.
- Also, age was divided into categories and analyzed based on profile photos.

Data Anonymization

- Anonymization of the last names of each individual commenter was a huge undertaking. We used the in-built library hashlib to transform last names into individual MD5 hexadecimal hashes. This ensures that in case there are two comments from the same person, their last name is not human-readable.
- Example:
- Novotný → anonymization MD5 function → a8ceff6567ebe7a54af101161b74f3f0

Vulgarisms

- We have written a function that uses regular expressions to look for vulgarisms in comments. The function would return number 1 if it found a match in the comment and 0 if there will be no match. We have included literal stars in the search, as some people try to censor insults with star signs

Data unification

- Since we were able to scrape a different number of comments under a different number of posts, it was important to unify the base of the data to proceed with the analysis.
- Firstly, we created a database with a minimal number of comments by each medium. The minimum number of comments was 1861 for TN.cz. 8 media by 1861 comments allowed us to work with 14 888 comments in total. The number of posts for a unified number of comments changed slightly.
- We also created a second database with the same number of posts for each medium; in our case, it was the minimal value of 27 posts by ČT24. Because some media, for example, iDnes or PrimaFTV have lower average of comments under 1 post, the quantity of comments dropped significantly, and we worked only with 7589 comments.

- All the visualizations of this research were created using the Microsoft Power BI database and a tool called Lynt.cz, which can count the frequency of words.
- Some of the data were also worked with the second one, based on the context, to get answers to some of our hypotheses.

Results

- • ČT24 has, in average, the most comments under 1 post, iDnes.cz or PrimaFTV have in average the least comments under 1 post
- • Seznam Zprávy and Aktuálně.cz have slightly more regular commentators, TN.cz and Blesk have fewer regular commentators.
- • Women comment more on TN.cz and Blesk, men comment more on iDnes.cz and Aktuálně.cz
- • Blesk and iDnes have more middle age readers, Aktuálně.cz, Seznam Zprávy, and TN are more popular among younger readers.
- • Blesk has the most readers from Jihočeský region, Prague comments largely on Seznam Zprávy.

- Highest number of fans commented on ČT24, Aktuálně.cz and TN.cz had no comments from top fans.
- People read and are active on multiple media platforms at the same time; they don't stick to only one medium.
- PrimaFTV has the most vulgar comments (8,4%).
- The most vulgar commentators
- Number of likes does not influence the amount of comments

Results - segmentation

- In Seznam Zprávy we can see that most commentators are men (68,57 %) in the youngest group (age 18-30) that are from our capital city Prague. Seznam Zprávy is a very young medium created in May 2016, so we can expect bigger expansion and more readers and followers on social media.
- IDnes.cz also has more men commentators (70,9 %), but the biggest age category is middle age 30-50 years from Prague. What is interesting is that they have very little commentators in the oldest age group even though it is a medium owned by our Czech prime minister who is voted mostly by older people.
- Aktuálně.cz was created at the same time as iDnes.cz but has fewer posts and less commentators. The segmentation here is mostly men (70,56 %). Interesting is that most commentators are young people from 18 to 30 (521 commentators), and they are mostly from the Zlínský region, not Prague as we would expect from the size of these regions.

- Novinky has quite a similar segmentation to iDnes.cz. Most commentators are men, with 66,77 % in the middle age group from 30 to 50 years from again capital city Prague. Facebook page Novinky.cz was also created in 2009, only a few months after iDnes.cz, and even though they have less posts, they have more unique commentators than iDnes.cz.
- ČT24 Facebook page was also created in 2009 as Novinky.cz, Aktuálně.cz and iDnes.cz. They have only 20 posts but an impressive number of 1342 commentators for such a small number of posts. Most of these commentators are again men (66,54 %). When it comes to location, the situation is the same as with most of the media, the highest number of commentators is from Prague.
- TN.cz Facebook page has more women commentators (57,62 %), which is not such domination as when it comes to other media with men, but it's still a significant amount. We could assume that it can be due to the content TN.cz has, but this page has similar posts to other media. TN.cz also has a lot of commentators, mostly young people from Prague.

- Zpravodajství FTV Prima or Prima News has an expected segmentation. The majority of commentators are men (53,16 %) from Prague. Age groups are almost balanced, slightly bigger is the middle age group. FTV Prima has a lot of posts, but the number of commentators is not that significant compared to other media.
- Blesk.cz is a medium commented mostly by women (66,71 %). We compare this to the fact that women like to gossip more than men and Blesk.cz is a tabloid that posts except for the local and foreign news about politics and economics and so also gossip about celebrities.

Conclusion

- We were able to acquire over 16 700 comments under almost 500 posts from 8 selected Czech media Facebook pages.
- Using this data, we analysed their reader base and its behaviour in the comment section.
- We found out what media are mostly read by different age groups or people from different regions.
- Finally, we managed to confirm 9 out of 12 hypotheses, which were based on common stereotypes.
  - We can say that some of the gender or age stereotypes are actually correct.
  - Our results denied the hypothesis that Women write longer comments because they are more talkative, Women comment more, or that young people read "only online media" more than older people, they read "traditional media".

Limitations and further work

- Currently we were hashing only the last name of the person - we did not create a unique identifier while acquiring the data, we created a unique identifier for each reader during the data processing, using a combination of their first name, last name, gender, age, and region. Creating a unique identifier using the link to their Facebook profile could be more relevant.
- acquire more data, such as data on education or employment of the commentators. But Facebook profiles of commentators are mostly private.
- It could be maybe possible for the media themselves because of their insight into the Facebook page statistics. We think that analysis of the content of comments is a very interesting field, and we can get a lot of interesting information and get into the minds of people. The media could use the analysis to find out what their readers are most interested in or what their opinions are on different topics. This kind of analysis could be beneficial also for companies that could discover what do people think about their products or services, but also for the governance, to discover what is the public opinion on different topics. This kind of analysis has the only and the biggest problem, the data.

Questions?