Temporal Evolution of Topics on Twitter

Daniel Pereira, Wladmir Brandão and Mark Song
Programa de Pós-Graduação em Informática, Instituto de Ciências Exatas e Informática, Pontifícia Universidade Católica de Minas Gerais, Brazil

Keywords: Topic Evolution, Twitter, Formal Concept Analysis, Social Network Analysis.

Abstract: Social networks became an environment where users express their feeling and share news in real-time. But analyzing the content produced by the users is not simple, considering the number of posts. It is worthy to understand what is being expressed by users to get insights about companies, public figures, and news. To the best of our knowledge, the state-of-the-art lacks proposing studies about how the topics discussed by social network users change over time. In this context, this work measure how topics discussed on Twitter vary over time. We used Formal Concept Analysis to measure how these topics were varying, considering the support and confidence metrics. We tested our solution on two case studies, first using the RepLab 2013 and second creating a database with tweets that discuss vaccines in Brazil. The result confirms that is possible to understand what Twitter users were discussing and how these topics changed over time. Our work benefits companies who want to analyze what users are discussing about them.

1 INTRODUCTION

The Internet is no longer just a repository for documents to be shared, it is now a hybrid space for different media and applications that reach a large audience (Zhang et al., 2012). Some of these applications are social networks, which allow their users to generate a large amount of content that exemplifies their impressions and experiences. A specific social network that stands out for forcing its users to express themselves concisely is Twitter. On Twitter, users express themselves through tweets, which consist of a text content with a maximum length of 280 characters.

The fact that the tweet is a short textual model allows users to quickly report what they are experiencing at the time the post is posted, unlike a journalist, for example, who, to generate a story, needs to ensure its excellence. Since Twitter users report their experiences without worrying about their writing or who will read their text, Twitter is probably the fastest means of disseminating information in the world (Cataldi et al., 2010).

With this large amount of information provided, it is hard to extract knowledge from a group of tweets. This task is relevant for companies, for example, to check the opinion that users are expressing about them. Therefore, the RepLab 2013 database was created, which groups tweets according to the subject addressed, grouping these tweets into entities, which can be companies, celebrities, or organizations. However, the entities are defined manually to ensure the assertiveness of the database (Amigó et al., 2013).

An alternative to solve this challenge is through Natural Language Processing (NLP) and Formal Concept Analysis (FCA). The objective of our work is to use NLP to find recurrent groups of words from tweets and then analyze how these groups of words relate to each other. The relation between these terms is measured using FCA, using the metrics of support and confidence. Also, how these terms change over time is another metric analyzed.

We worked on two case studies to solve the problem. The first case study used RepLab 2013 to analyze the BMW entity and then create a new database using Twitter API with the term BMW as a query. The second case study consists of analyzing tweets that discuss vaccines in Brazil in January 2021, checking which terms are related to vaccines and how they evolve over time.

The remainder of the paper is organized as follows: the background is outlined in Section 2. The Literature Review is described in Section 3. Section 4 presents the defined Methodology. Results are discussed in Section 5. The conclusion and further research are in Section 6.

Pereira, D., Brandão, W. and Song, M.
Temporal Evolution of Topics on Twitter.
DOI: 10.5220/0011524500003318
In *Proceedings of the 18th International Conference on Web Information Systems and Technologies (WEBIST 2022)*, pages 113-119
ISBN: 978-989-758-613-2; ISSN: 2184-3252

2 BACKGROUND

2.1 Formal Concept Analysis

FCA is a technique based on formalizing the notion of concept and structuring concepts in a conceptual hierarchy. FCA relies on lattice theory to structure formal concepts and enable data analysis. The capability to hierarchize concepts extracted from data turns FCA an interesting tool for dependency analysis. With the increase of social networks and due to the large amount of data generated by users, the study and improvement of techniques to extract knowledge are becoming increasingly justified. Also, it permits the data analysis through associations and dependencies attributes and objects, formally described, from a dataset.

2.1.1 Formal Context

Formally, a formal context is formed by a triple (G,M,I), where G is a set of objects (rows), M is a set of attributes (columns) and I is defined as the binary relationship (incidence relation) between objects and their attributes where $I \subseteq G \times M$.

Table 1 exemplifies a formal context. In this example, objects correspond to tweets, attributes are the characteristics (terms), and the relationship of incidence represents whether or not the tweet has that characteristic. A tweet has that characteristic if and only if there is an $'X'$ at the intersection between the row and the respective column.

Table 1: Formal Context Example.

	Used BMW	Pay Online	BMW X5	BMW M3
Tweet 1	**X**			
Tweet 2		**X**		**X**
Tweet 3	**X**	**X**		**X**
Tweet 4			**X**	

2.2 Formal Concepts

Let (G,M,I) be a formal context, $A \subseteq G$ a subset of objects and $B \subseteq M$ a subset of attributes. Formal concepts are defined by a pair (A,B) where $A \subseteq G$ is called extension and $B \subseteq M$ is called intention. This pair must follow the conditions where $A = B'$ and $B = A'$ (Ganter and Wille, 1999). The relation is defined by the derivation operator ($'$):

$$A' = \{\, m \in M |\ \forall\, g \in A, (g, m) \in I\}$$
$$B' = \{\, g \in G |\ \forall\, m \in B, (g, m) \in I\}$$

If $A \subseteq G$, then A' is a set of attributes common to the objects of A. The derivation operator ($'$) can be reapplied in A' resulting in a set of objects again (A''). Intuitively, A'' returns the set of all objects that have in common the attributes of A'; note that $A \subseteq A''$. The operator is similarly defined for the attribute set. If $B \subseteq M$, then B' returns the set of objects that have the attributes of B in common. Thus, B'' returns the set of attributes common to all objects that have the attributes of B in common; consequently, $B \subseteq B''$.

As an example, using Table 1, objects A $= \{Tweet2, Tweet3\}$, when submitted to the operator described above, will result in $A' = \{PayOnline, BMWM3\}$. So $\{\{Tweet2, Tweet3\}, \{PayOnline, BMWM3\}\}$ is a concept. All concepts found from Table 1 are displayed in Table 2.

Table 2: Existing concepts in the formal context of Table 1.

Objects	Attributes
{Tweet 1, Tweet 2, Tweet 3, Tweet 4}	{}
{Tweet 4}	{BMW X5}
{Tweet 1, Tweet 3}	{Used BMW}
{Tweet 2, Tweet 3}	{Pay Online, BMW M3}
{}	{Used BMW, Pay Online, BMW X5, BMW M3}

In Table 2 there is a concept with an empty attribute set and a concept with an empty object set. They are called *infimum* and *supremum*, respectively.

2.2.1 Implication Rules

Implications are dependencies between elements of a set obtained from a formal context. Given the context (G,M,I) the rules of implication are of the form $B \rightarrow C$ if and only if $B, C \subset M$ and $B' \subset C'$ (Ganter et al., 2005). An implication rule $B \rightarrow C$ is considered valid if and only if every object that has the attributes of B will also have the attributes of C.

We can define rules, as follows: $r : A \rightarrow B(s,c)$, where $A, B \subseteq M$ and $A \cap B = \varnothing$. We can also define the support of the rules, which is defined by $s = supp(r) = \frac{|A' \cap B'|}{|G|}$ and the confidence of the rules, which is defined by $c = conf(r) = \frac{|A' \cap B'|}{|A'|}$(Agrawal and Srikant, 1994).

Table 3 shows two existing rules in the context of Table 1. The rule Pay Online $\rightarrow$ BMW M3 has 50% support because this rule happens in 2 tweets, out of a total of 4 tweets. The confidence is 100%, since whenever a tweet has Pay Online it also has BMW M3. When a rule has 100% confidence, such as the rule Pay Online $\rightarrow$ BMW M3, it is called an implication.

Table 3: Example of supported and trusted rules.

Rule	Support	Confidence
Pay Online → BMW M3	50%	100%
Used BMW → Pay Online, BMW M3	25%	50%

2.3 Database Processing

Textual databases, such as RepLab 2013, need to be pre-processed before being analyzed. The steps performed in this work are the following: N-Gram, stop word removal and Regular Expression.

- N-Gram: is a contiguous sequence of n items from a given sample of text. The items can be letters or words that are in sequence on a text sample;
- Stop word removal: consists of removing words such as articles and prepositions, as these words are not significant for textual analysis;
- Regular Expression: a technique to determine a pattern in a text sample. It is used to find a group of words that need to be replaced or deleted.

The steps described above were applied through the Python package Natural Language Toolkit (NLTK). The NLTK package has a list of stop words, such as “the”, “a”, “an”, “in”, so those words in the list are removed from the database being preprocessed, as these words are not meaningful to the analysis. This allows the database after pre-processing to have a reduced size and also reduces the analysis time (Contreras et al., 2018).

An n-gram is a contiguous sequence of n items from a given sample of text. The items can be letters or words that are in sequence on a text sample. An n-gram of size 1 is referred to as a unigram and does not consider other words that are in sequence. Size 2 is a bigram and size 3 is a trigram meaning that a group of three words are in sequence in a text sample (Roark et al., 2007). Table 4 shows an example of bigrams and trigrams found on a text sample.

3 LITERATURE REVIEW

Several works are relevant to the context of this study. These are works in the context of topic detection in social networks, topic evolution, and classification of textual bodies. These works are described in the next paragraphs.

Zhang et al. (Zhang et al., 2012) detail how the detection of topics on the Internet is a challenge because the information produced on the Internet is succinct and does not adequately describe the real context being addressed. To solve this characteristic of the information produced on the Internet, the authors used the technique pseudo-relevance feedback, which consists of adding information to the data being analyzed.

With this strategy, the authors were able to improve the information produced on the Internet, improving the context that this information is dealing with and thus being able to identify within this information which will become more present on the Internet in the future. This research also seeks to detect topics of content produced on Twitter but we did not use the pseudo-relevance feedback technique, since the RepLab 2013 database already provides us with the context in which the analyzed content belongs.

Cataldi et al. (Cataldi et al., 2010) used the topic detection technique to identify emerging topics in the Twitter community. The authors were able to carry out the identification considering that if the topic occurs frequently in the present and was rare in the past, and thus characterized them as emerging. To enhance the strategy addressed, an analysis of the authors of these emerging topics was carried out through the Page Rank algorithm, to ensure that the emerging topic is not present only in some bubble of the Twitter community. Finally, a graph was created that connects the emerging topic with other topics that are related to it, and that therefore have a greater chance of becoming emerging topics as well. Unlike the work described above, this research aims to use topic detection to analyze how these topics change over time.

Dragoş et al. (SM. et al., 2017) present an approach that investigates the behavior of users of a learning platform using FCA. The log generated by the platform contains information about the actions that each student is performing on the platform. So the log allows to identify the profile of students.

The use of FCA by Dragoş et al. occurs to consider the instant of time that the actions are performed by the students. It is relevant to profile students to understand whether they are performing actions late, early or on time. Therefore, FCA can be considered as an alternative to study temporal events.

Cigarrán et al. (Cigarrán et al., 2016) used FCA to group tweets according to the topics found. That’s why the RepLab 2013 database was used, which already groups tweets into entities, based on the textual content of the tweet. By using FCA, the work still manages to obtain a conceptual grid of the topics found, obtaining a hierarchical view of the topics, which is a differential in relation to other techniques. The proposal was among the best results of the RepLab 2013 forum, proving the effectiveness of FCA for the topic detection challenge.

Table 4: Example of bigram and trigram.

Text Sample	Bigram	Trigram
Topics change over time	{Topics change} {change over} {over time}	{Topics change over} {change over time}

Amigó et al. (Amigó et al., 2013) describes the organization and results of RepLab 2013, which focuses on monitoring the reputation of companies and individuals through the opinion of Twitter users. This is done by dividing the tweets into entities, with each entity comprising a company or an individual. Within the entities, it is evaluated whether the tweet presents positive or negative aspects to the entity. In this work it will not be observed whether the tweets have a positive or negative aspect to the entity, the focus will be on detecting topics present in the tweets and how they vary over time.

Arca et al. (Arca et al., 2020) propose an approach to suggest tags (meaningful human-friendly words) for videos that consider hot trend subjects, so the video will receive more access since it will be related to a trending subject. The original tags are inserted manually and these tags are the input for the algorithm, that will match them with a hot trend subject. Our proposed method also identifies meaningful words, the difference is that our input are tweets, and then analyzes how these words vary over time.

4 METHODOLOGY

This section presents our methodology to achieve the proposed objectives. For this, the steps presented in the sections below were performed.

4.1 RepLab 2013

The Replab Evaluation Campaign 2013 is an international forum for experimentation and evaluation in the field of Online Reputation Management. One of the challenges addressed in the forum is the classification of tweets into entities, which identify the topics that the tweet addresses.

The Replab 2013 database consists of a group of tweets related to 61 entities that were extracted between June 1, 2012 and December 31, 2012. These entities are divided into four domains, namely: automobiles, financial entities, universities and music/artists.

This database was chosen since the works (Amigó et al., 2013; Castellanos et al., 2017; Cigarrán et al., 2016) that address topic detection use RepLab 2013 to validate the proposed models and also because ReBLab has labeled the tweets into topics, topics that define the context of a tweet. This process of assigning labels to tweets was done manually and did not consider when the tweet was posted.

In this work, the proposed methodology is to use FCA to identify how topics found on tweets vary over time.

4.1.1 Treatment of RepLab 2013

To obtain all the information necessary to carry out the work an integration was made with the Twitter API to retrieve the text body and the publication dates of the tweets present in Rep Lab 2013. Rep Lab 2013 does not have this information to respect the privacy of the authors of the tweets, because if they delete the tweets it will prevent their post from being used by works that use Rep Lab 2013.

With the integration performed, it was possible to retrieve 32402 tweets and the posting date. With all the necessary information obtained, the next step is to pre-process the database, so that the textual body of the tweets is transformed into a list of words that will be analyzed. We chose to pre-process tweets from BMW entity, that belongs to automobiles domain, to analyze what authors were saying about BMW and how this changed over time. BMW entity has 942 tweets.

To perform the task, the techniques described in Section 2.3 were used, which were applied to the integrated database of Rep Lab 2013. Just below is a pseudo-code describing the process:

```
Begin
    ApplyNgramFunction();
    OrderNgramByFrequency();
    SelectMeaningfulNgram();
    CreateJsonFileForLatticeMiner();
    RunLatticeMiner();
    ExtractRules();
End.
```

N-Gram provided unigrams, bigrams and trigrams and the ones that described the context of a tweet were selected to be analyzed. The stop word removal method was used to remove unigrams that matched with a stop word. Finally, a regular expression removed the URLs, since a URL do not describe the context of a tweet.

To improve our work we decided to create a new database with recent tweets about BMW. For that we chose the 10 most frequent N-Grams found on the BMW entity from RepLab 2013 to be the query parameter on Twitter API. The 10 most frequently N-Grams are the following:

- BMW Series;
- M6 Gran Coupe;
- New BMW;
- BMW M3;
- BMW X5;
- BMW Z4;
- For Sale;
- Youtube Video;
- For Free;
- BMW M5.

For 8 days we used the Twitter API to get these new tweets. In the end, a different preprocessing method was used to remove the retweets. Retweets are tweets that have the same text content so we had to remove them to avoid implication rules over the same text context. With unique tweets, we used the techniques described in Section 2.3. After the preprocessing, the database has 3897 unique tweets.

4.2 Vaccines

In January of 2021, Brazil started to vaccinate against COVID-19. There was many discussions about this subject, if the vaccine was safe, government scandals about denying vaccine offers, and the lack of syringes to apply the vaccines. For that reason, we collected tweets over 13 days in January to check which terms were related to vaccines on Twitter.

Using Twitter API to search for tweets with the query vaccines, 105 tweets were collected. We searched for tweets with high engagement since these tweets represent the opinion of a large group, letting us extract the most commented terms about vaccines from a small group of tweets (Miao et al., 2016). Twitter API defines if a tweet has a high engagement and provides a parameter to filter these tweets.

4.3 Applying FCA

The selected N-Grams were analyzed by the Lattice Miner tool (Kwuida et al., 2010). We designed a formal context with the N-Grams and the creation date of a tweet to be the input of Lattice Miner. The formal context has the tweet identification as an object, N-Grams as attributes, and the creation date as the binary relationship, being formalized in a JSON file. Lattice Miner displays the formal context as a table and uses an X to relate tweets to N-Grams and the date. Lattice Miner's output is a group of implication rules, showing how N-Grams relate to each other over time.

We chose the significant rules that bring some insight to the companies or organizations that are being analyzed, in this case, is BMW. We ordered the rules by chronological order and then checked if the implication rules were changing day by day.

5 RESULTS

5.1 BMW

The first result obtained is the analysis of the topic "BMW vehicles for sale" from BMW entity. Using tweets from RepLab 2013 we checked which BMW car models were being announced for sale on Twitter. The result is in Table 5.

Table 5: For Sale topic changing over time.

Day	Antecedent	Consequence	Support	Confidence
1	BMW Series	For Sale	0.65%	50%
1	BMW M3	For Sale	1.3%	50%
2	BMW Series	For Sale	0.65%	20%
2	BMW X5	For Sale	0.65%	50%
2	BMW Z4	For Sale	0.65%	100%
3	BMW M3	For Sale	1.3%	33%
3	BMW Z4	For Sale	0.65%	50%
3	BMW X6	For Sale	0.65%	100%
4	BMW Series	For Sale	1.95%	50%

These first results show that the models announced on Twitter to be sold change each day. So the BMW company can analyze that information to understand which models are more frequent in the second-hand market. These results show 4 days of data, but analyzing a longer period could bring even more relevant information to the BMW company.

Then we analyzed 942 tweets from the BMW entity, which belongs to RepLab 2013, without specifying any topic. The found implication rules were inside an interval of 5 days, between 2012-06-01 and 2012-06-05. The result is in Table 6.

These results confirm that the For Sale topic is relevant even analyzing the whole entity. Another observed pattern is that during 5 days Twitter's users talked about BMW and Audi but on days 4 and 5 they started to publish about Mercedes too. With that information, BMW company could investigate better to understand why Twitter users are talking about these

Table 6: BMW entity implication rules.

Day	Antecedent	Consequence	Support	Confidence
1	BMW	Audi	0.46%	1.92%
1	BMW	Buy	1.39%	5.76%
1	BMW	For Sale	0.46%	1.92%
2	BMW	Audi	0.93%	3.27%
2	BMW	For Sale	0.46%	1.63%
2	BMW	Buy	0.93%	3.27%
3	BMW	Audi	0.46%	3.99%
4	BMW	Audi, Mercedes	0.46%	4.34%
4	BMW	Want	0.93%	8.69%
5	BMW	Audi	3.72%	22.22%
5	BMW	Mercedes	2.79%	16.66%

car brands.

At last, we analyzed 3897 tweets that were collected by us to see the results from recent tweets, since RepLab 2013 tweets were collected in 2012. The found implication rules were inside an interval of 5 days. The result is in Table 7.

Table 7: BMW entity implication rules from tweets collected by us.

Day	Antecedent	Consequence	Support	Confidence
1	Used BMW	Pay Online	0.36%	4.99%
2	BMW M4 CSL	Passion and Confidence	1.09%	8.33%
4	Used BMW	Pay Online	0.36%	16.66%
5	Used BMW	Pay Online	0.36%	50%

The results show that used BMW cars are related to online payment, showing that there is an advance in this market. We also could identify that BMW M4 CSL, a new BMW car, relates to Passion and Confidence, the slogan of an eSports tournament that BMW sponsors. That information shows the company which car model is being affected by their sponsorship of the tournament, revealing if the target of this marketing was accomplished.

5.2 Vaccine

Table 8 shows the obtained results from the tweets that discuss vaccines. Day 1 represents January 5th, 2021.

Analyzing the tweets related to vaccines in Brazil we realized that Brazilian president Bolsonaro was mentioned in tweets that discuss vaccines almost every day we analyzed. This relationship makes sense since Bolsonaro is against COVID vaccines and did several public speeches to discourage Brazilians from getting vaccinated.

Bolsonaro also got related to syringes on the second day of our analysis. That happened because the Brazilian government did not provide enough syringes to start the vaccination process. This rule only appeared on day 2, showing how volatile the discussions on Twitter are.

The efficiency of vaccines was discussed through the days we analyzed since a rule linking vaccines with efficiency happened on four different days. An explanation for the continuity of efficiency discussion is that 4 different COVID vaccines are used in Brazil, so Twitter users discuss the efficiency of each one on a specific day.

Another aspect is that the rules Vaccine implicates in Bolsonaro and Vaccine implicates in Efficiency have a relation between each other that if one appears in one day the other one does not, or has a low support. An explanation is that Brazilian president can not interfere on vaccines' efficiencies, so the support of the rule Vaccine implicates on Bolsonaro is higher on days that problems caused by Brazilian government happened, like day 10 and 11, when an oxygen crisis in the Brazilian city Manaus was neglected by Bolsonaro's government.

These results show that our approach matches with the news about vaccines from January 2021 and brings insights like the negative correlation between the rules of Bolsonaro and efficiency. Using it during other months and different subjects can also provide good results.

Table 8: Vacine implication rules.

Day	Antecedent	Consequence	Support	Confidence
1	Vaccine	Bolsonaro	1%	50%
2	Vaccine	Bolsonaro	4%	41%
2	Syringes	Bolsonaro	2%	75%
3	Vaccine	Efficiency	4%	41%
7	Vaccine	China	6%	40%
8	Vaccine	Efficiency	5%	50%
8	Vaccine	Bolsonaro	1%	16%
9	Vaccine	Efficiency	3%	57%
9	Vaccine	Bolsonaro	1%	28%
10	Vaccine	Manaus	3%	40%
10	Vaccine	Efficiency	1%	20%
11	Vaccine	Oxygen	6%	38%
11	Vaccine	Bolsonaro	6%	38%
13	Vaccine	Bolsonaro	1%	14%

6 CONCLUSIONS

In this paper, we have proposed a technique to identify how topics discussed on Twitter change over time. To achieve that we used FCA to build contexts with the analyzed tweets and extract implication rules from these contexts. The metrics support and confidence are essential to measuring how these topics vary. To test the technique we used the RepLab 2013 database to provide tweets already divided into entities and also a database created for this paper with tweets from 2022. The results show that is possible to identify relevant topics for companies and how these topics change over time.

We would like to explore more tweets with a bigger time range to have a better view of emerging topics on Twitter and for how long they stay relevant. Also, a technique that can analyze tweets in real-time would be interesting to provide information to the company that is being analyzed at the same moment that the users are talking about a topic. These could improve the actions that a company could take about what is being said about it.

As future work we plan to reproduce the methodology during the Brazilian election period to understand what Twitter users are discussing about the candidates. It is a great opportunity since elections are a subject widely discussed, providing us a big quantity of tweets.

REFERENCES

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Amigó, E., Carrillo-de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., Rijke, M., and Spina, D. (2013). Overview of replab 2013: Evaluating online reputation monitoring systems. volume 1179.

Arca, A., Carta, S., Giuliani, A., Stanciu, M., and Reforgiato Recupero, D. (2020). Automated tag enrichment by semantically related trends. pages 183–193.

Castellanos, A., Cigarrán, J., and García-Serrano, A. (2017). Formal concept analysis for topic detection: A clustering quality experimental analysis. *Information Systems*, 66:24–42.

Cataldi, M., Di Caro, L., and Schifanella, C. (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, New York, NY, USA. Association for Computing Machinery.

Cigarrán, J., Ángel Castellanos, and García-Serrano, A. (2016). A step forward for topic detection in twitter: An fca-based approach. *Expert Systems with Applications*, 57:21–36.

Contreras, J. O., Hilles, S., and Abubakar, Z. B. (2018). Automated essay scoring with ontology based on text mining and nltk tools. In *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, pages 1–6.

Ganter, B., Stumme, G., and Wille, R., editors (2005). *Formal Concept Analysis: Foundations and Applications*. Springer.

Ganter, B. and Wille, R. (1999). *Formal concept analysis: mathematical foundations*. Springer, Berlin; New York.

Kwuida, L., Missaoui, R., Amor, B. B., Boumedjout, L., and Vaillancourt, J. (2010). Restrictions on concept lattices for pattern management. In Kryszkiewicz, M. and Obiedkov, S. A., editors, *Proceedings of the 7th International Conference on Concept Lattices and Their Applications, Sevilla, Spain, October 19-21, 2010*, volume 672 of *CEUR Workshop Proceedings*, pages 235–246. CEUR-WS.org.

Miao, Z., Chen, K., Fang, Y., He, J., Zhou, Y., Zhang, W., and Zha, H. (2016). Cost-effective online trending topic detection and popularity prediction in microblogging. *ACM Trans. Inf. Syst.*, 35(3).

Roark, B., Saraclar, M., and Collins, M. (2007). Discriminative n-gram language modeling. *Computer Speech Language*, 21(2):373–392.

SM., D., C., S., and Şotropa DF. (2017). An investigation of user behavior in educational platforms using temporal concept analysis. In *ICFCA 2017*.

Zhang, J., Liu, D., Ong, K.-L., Li, Z., and Li, M. (2012). Detecting topic labels for tweets by matching features from pseudo-relevance feedback. In *Proceedings of the Tenth Australasian Data Mining Conference - Volume 134*, AusDM '12, page 9–19, AUS. Australian Computer Society, Inc.