Agris on-line Papers in Economics and Informatics

Volume XI Number 4, 2019

Measuring the Similarities of Twitter Hashtags for Agriculture in the Czech Language

John Phillip Sabou[1], Petr Cihelka[1], Miloš Ulman[1], Dana Klimešová[2]

[1] Department of Information Technology, Faculty of Economics and Management, Czech University of Life Sciences, Prague, Czech Republic

[2] Department of Information Engineering, Faculty of Economics and Management, Czech University of Life Sciences in Prague, Czech Republic

Abstract

Our paper presents first analysis of Czech Twitter content within the agriculture context. We deployed textual analysis of more than 240,000 tweets over 2014-2019 hashtags that were, according to Google Trends, most trending and related to Czech agriculture such as #dotace, #repka, or #bionafta – both in Czech and English language. Besides descriptive statistics of the tweet dataset, we visualized keyword correlations which revealed strong focus of the discourse on rapeseed, biofuel and the prime minister Andrej Babiš. Owing to inherent political context of the given hashtags, we found spikes in topics which followed the public attention to the topics in mass media. We also found several accounts that produces high traffic for certain hashtags in Czech, yet those accounts were located abroad. Consistent with other studies, a high proportion of tweets was generated by unverified accounts that might be bots – automated accounts. We propose to conduct semantic analysis of a broader dataset over the main social media platforms in the Czech Republic.

Keywords

Agriculture, Twitter, Czech language, word occurrence, descriptive statistics.

Sabou, J. P., Cihelka, P., Ulman, M. and Klimešová, D. (2019) "Measuring the Similarities of Twitter Hashtags for Agriculture in the Czech Language ", *AGRIS on-line Papers in Economics and Informatics*, Vol. 11, No. 4, pp. 105-112. ISSN 1804-1930. DOI 10.7160/aol.2019.110410.

Introduction

This paper explores the semantic similarity of Czech Twitter messages by using Descriptive Statistics to calculate the number of times a hashtag is repeated in a Twitter corpus in the Czech language. These are classically reduced to their frequency of co-occurrence in language: the more frequently two words appears together, the higher is their similarity. The goal of this paper is to discern which topics in the Czech language seem to be artificially diffused via Twitter. We will do this by computing the similarity and co-occurrence data in a large corpus using descriptive statistical analysis.

Online news is a well-researched field, with many good reasons why artificial intelligence (AI) techniques have the potential to improve the way we consume news online (Orhan, 2017; Mazurek et al., 2019). At the same time, news is a biased form of media that is increasingly driven by content that can sell advertising. Some stories that may be of interest often get buried, while other content may receive greater exposure in seemingly artificial ways. For example, Google News is a topically segregated mashup of feeds, with automatic ranking strategies based on user interactions (click-histories).

There is considerable research attention being paid to Twitter and the internet in general. These services provide access to new types of information and the real-time nature of these data streams provide as many opportunities as they do challenges. In addition, companies like Twitter have adopted a very open approach to making their data available and Twitter's developer API provides researchers with access to a huge volume of information.

Studies on large corpora have given examples of words that have strong associations with one another, although they never co-occur in paragraphs. For instance, Lund & Burgess (1996) mentioned the two words road and street that almost never cooccur in together. However, both words

are strongly associated. The correlation between co-occurrence and similarity has been found by several researchers (Spence and Owens, 1990). This relation can be viewed as a simplification of Miller and Charles' (1991) hypothesis: "Two words are semantically similar to the extent that their contextual representations are identical".

Czech Agricultural Sector

The emergence of social media has generated renewed attention to rhetoric via online tweets, blogs, and comments on public issues. Social media users are not demographically representative and diverse social media platforms undoubtedly develop local cultures of expressive style that influence the character of what people choose to say. Nonetheless digital contents are an important, instantiation of public opinion (Watrobski et al., 2016). We are interested in how this instantiation plays out with the Czech Agricultural Sector.

Researchers at the Czech University of Life Sciences have been forming a multifunctional understanding of the Czech agricultural sector and its relationship to the EU Common Agricultural Policy, which is focused on keeping farmers in rural areas with cost-efficient technology adaptations ((Svatoš and Smutka, 2009; Vaněk et al., 2010; Kołodziejczak and Kossowski, 2011; Věžník, Král, and Svobodová, 2013; Reiff et al., 2016). In this respect, it is important to understand the underlying issues that form around Czech agricultural issues. That includes any influences, whether internal or external, that affect how the various stakeholders of the sector interact with one another and engage the public in general.

Problem Statement

We explore whether Twitter hashtags have a high order of co-occurrences, and whether that plays an important role in the construction of word similarities in the Czech language via Twitter. The work of this paper follows on previous research regarding the flow of information on platforms such as Facebook and Twitter. These dealt with how information proliferates through networks and how topics artificially dominate the discourse space (Wald et al., 2013; Xu et al., 2011; Ozdikis et al., 2012; Jansen et al., 2009). Our key questions include:

RQ1: Which hashtags occur the most frequently in the Czech agricultural sector?

RQ2: What proportion of Twitter accounts are verified versus non-verified in the Czech agricultural discourse?

Materials and methods

Descriptive statistics are brief coefficients that summarize a given data set, which can be either a representation of the entire or a sample of a population. All descriptive statistics are either measures of central tendency or measures of variability, also known as measures of dispersion. For our analytical framework we will measure the spread or dispersion of the data points.

Estimating Word Occurrences

Turney (2001) defines a method for estimating word similarity based on Church and Hanks' (1990) pointwise mutual information. The mutual information between x and y is defined as the comparison between the probability of observing x and y together and observing them independently:

$$I(x,y) = \log(p(x,y) / p(x).p(y)).$$

By extension, this model provides a way to measure the degree of co-occurrence of two words by comparing the number of co-occurrences to the number of individual occurrences (Bordag, 2018). In our case, a Twitter hashtag in the Czech language is accompanied with similar hashtags that supposedly have different meanings, yet lead to the same information, then we may deduce what topics users are clustering around related to Czech agriculture.

A central question in text mining and natural language processing is how to quantify what a document is about. One measure of how important a word may be is its term frequency (*tf*), how frequently a word occurs in a document. Another approach is to look at a term's inverse document frequency (*idf*), which decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents (Chaudari et al., 2011). We can use this approach to the analysis to quantify how important various terms are in a document that is part of a collection.

Input Criteria

The Twitter dataset was collected from an archive of hashtags between December 1st, 2014 to November 1st, 2019. All tweets had at least one of the selected hashtags: #babis, #agrofert, #capihnizdo, #dotace, #repka. The English equivalents of those selected hashtags were also used. In total, 10 hashtags were monitored and 194,716 tweets were retrieved from Twitter.

Hashtags	Total count of tweets retriever
#babis	51,348
#agrofert	6,677
#bionafta #biodiesel	71,110
#capihnizdo	4,774
#dotace	36,971
#repka #rapeseed	23,836

Source: Own processing

Table 1: Number of Tweets for each hashtag.

After the data retrieval, a descriptive analysis was performed on the dataset using an in-house developed algorithm. Each dataset for the selected hashtags were compiled and analyzed according to how many tweets was published, favorited, quoted, and replied. In this way we could see how many individual, public accounts published the tweets versus verified accounts. We also achieved an overview of these accounts according to the number of published tweets, favorited tweets, retweets, replies and quotes. These were also compiled according to the accounts' *country, city, language* and *user* origin.

After the descriptive analysis, a textual analysis was performed on the text of the tweets and hashtags. The target of this analysis was to show a relationship between used keywords and correlations between the terms used in the tweets. This was performed with an analysis of the hashtags followed by an analysis of the keywords used from the whole dataset that was extracted from Twitter.

Results and discussion

Analysis of the hashtags consisted of a detection threshold for published tweets and the reply by users via another tweet, a retweet or a quoted tweet by someone else if it was favorited. For each of these metrics we were able to get a count. If a tweet was in reply to another tweet, then an analysis of that relationship was given. However, due to limitations of the data set concerning the timeframe collection (2014-2019), we were not always able to find the original tweet linked to the hashtag if it was not already in the dataset. Table 2 shows us the number of hashtags that appeared to have moderate to high levels of word occurrence and cooccurrence. Table 3 augments this information with the total number of accounts and their tweets.

From the collected dataset we were able to find out how many accounts were validated, and how many were not. We can assume that the verified accounts should be more trustworthy than the accounts that are not verified. There is also the possibility that non-verified account could be bots, however we did not have the means to detect this. But in our case, it does mean that there was a disproportionate number of tweets that came from unverified users rather than verified users, indicating a behavior of anonymity in the activity of reusing tweets, e.g. #babis or #bionafta. This is not necessarily uncommon, as most Twitter users are unverified. However, it is unusual in this case considering that the tweets are specifically about the Czech

Hashtag	Total	Quoted	Favorited	Retweeted	Replied
#agrofert	6,677	147	1,611	761	665
#babis	51,348	1,459	10,492	5,584	4,309
#bionafta	71,110	1,114	12,317	11,986	2,047
#capihnizdo	4,774	191	1,168	580	504
#dotace	36,971	604	6,022	5,120	2,069
#repka	23,836	417	4,475	3,348	1,060

Source: Own processing

Table 2: Number of hashtags that cooccur with other topics.

Hashtag	Total number of accounts	Verified accounts	Tweets by verified accounts
#agrofert	2,746	50	218
#babis	16,464	366	1,706
#bionafta	21,912	619	1,821
#capihnizdo	1,922	23	105
#dotace	21,573	601	1,719
#repka	9,962	139	384

Source: Own processing

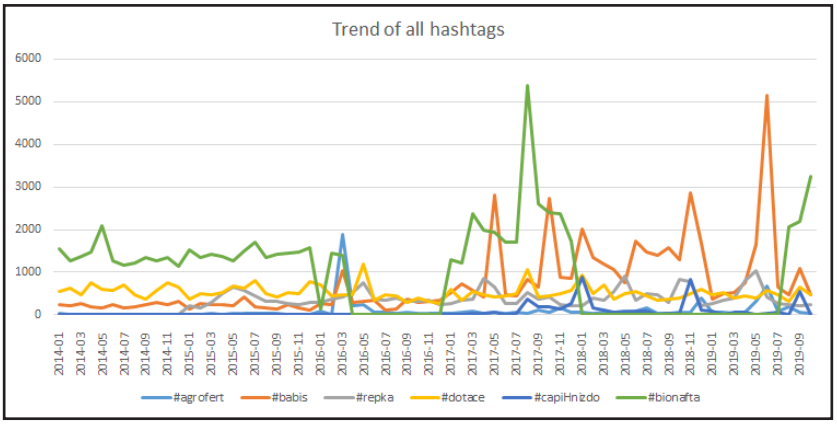
Table 3: Total number of accounts, verified accounts and tweets.

agricultural issues whose contexts are inherently political topics.

There is a strong indication that at certain points between 2014 – 2019, there were significant spikes in reused Twitter material in support of certain hashtags. Figure 1 shows the trends of each hashtag, where we can see spikes in public interest via Twitter for the topics such as #babis #bionafta and #dotace. The largest spike was between July and October 2017, which can be attributed to the parliamentary elections in October 2017. At that time, the political party ANO won with Andrej Babiš as their chairman, and he was appointed as the Prime Minister (Pehe, 2018).

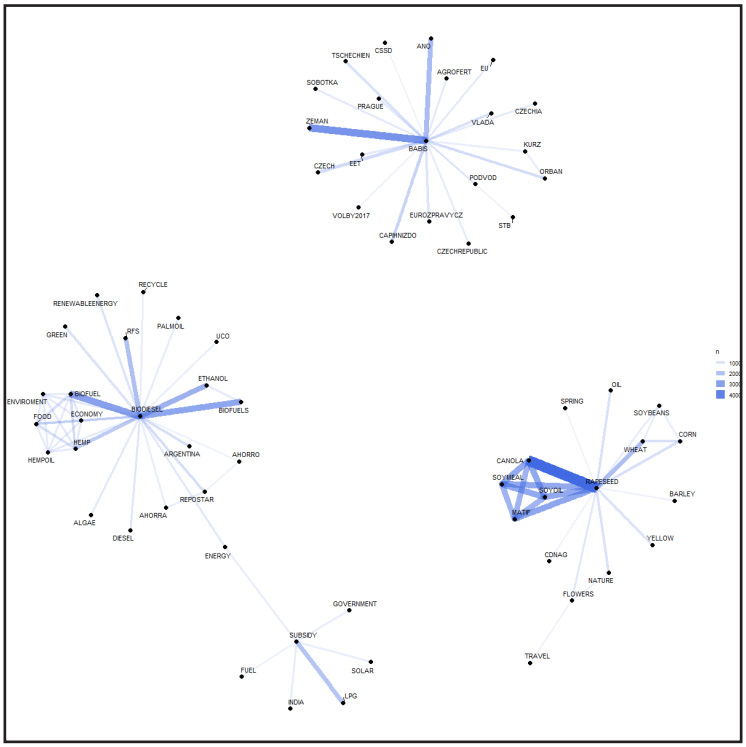
After the descriptive analysis of these, a textual analysis was performed on the content of the tweets. The target of this analysis was to show a relationship between used keywords and to find out whether there was a correlation between the terms used in the tweets. This was done by performing an analysis of the keyword pairs that commonly occur most often. All keywords were converted to uppercase to eliminate difference between letter casing, as well as removing diacritics and accents.

Figure 2 offers a map of these relationships in which 184,834 keywords show strong



Source: Own processing

Figure 1: Trends per hashtag between 2014-2019



Source: Own processing

Figure 2: Map of keyword correlations.

correlations that are used in Twitter hashtags. These keywords essentially denote the most popular topics concerning Czech agriculture around which users cluster. The blue lines indicate the strength of the relationship, whereby users are pushing those hashtags to bring attention to the topic(s). In this case, the strongest relationship is apparent between the keyword's "rapeseed", "canola" and "soymeal". For the graphic output we displayed only keyword pairs with occurrence over 500 counts. Detailed figures on the top 20 keyword pair occurrences are presented in Table 4.

#	Keyword pair	No of occurrences
1	RAPESEED - CANOLA	4,731
2	BABIS - ZEMAN	3,381
3	BIOFUEL - BIODIESEL	2,962
4	RAPESEED - MATIF	2,822
5	BIOFUELS - BIODIESEL	2,811
6	SOYOIL - SOYMEAL	2,750
7	RAPESEED - SOYMEAL	2,725
8	RAPESEED - SOYOIL	2,724
9	CANOLA - SOYOIL	2,723
10	CANOLA - SOYMEAL	2,717
11	BIODIESEL - ETHANOL	2,475
12	MATIF - SOYMEAL	2,386
13	CANOLA - MATIF	2,385
14	SOYOIL - MATIF	2,385
15	ANO - BABIS	2,133
16	RAPESEED - WHEAT	2,033
17	BIODIESEL - RFS	2,030
18	LPG - SUBSIDY	1,882
19	BIODIESEL - HEMP	1,683
20	BABIS - CAPIHNIZDO	1,419

Source: Own processing

Table 4: Top 20 keyword pair occurrences.

In contrast to the past work there has been interest in the detection of fake users from both online social networks and computer networking communities (Dickerson et al., 2014). The openness of Twitter's platform allows for, and even promotes, programs called "bots" that automatically post content. These bots post content ranging from helpful (e.g." recent news stories or public service announcements) to malicious spam or phishing links. Such bots on Twitter have become a nuisance, even triggering a long diatribe, particularly around political figures such as Andrej Babiš.

Similarly, Campoy (2019) reported that 60%, 33 million followers on Twitter were suspiciously inactive for longer than 120 days following his election in 2016. Yet these same accounts have been observed to become active during highly publicized issues important to his administration. In short, there is now a widespread belief that bots constitute a significant part of the social media world - and that many of them can be identified by the frequency of how often they reuse content through word occurrence. While this is only a piece of the puzzle, measuring word occurrence allows a glimpse of what is happening at a surface level. A further semantic analysis of the actual tweets would be needed to triangulate the veracity of suspicious accounts.

In relation to our results, it would seem that there was significant public discussion around rapeseed, which is a principal agricultural export in the Czech Republic (Carré, 2014). Social media represents one of the most important informational mediums that rural communities in the Czech Republic utilize to engage in public discussions (Červenková et al., 2008). It is not a surprise that rapeseed production, especially if subsidized by the EU Commission, would account for significant Twitter activity. Our analysis shows which keywords are more likely to occur together with other topics in Twitter concerning Czech agriculture.

What is unusual, is that we also found that several accounts producing high-level traffic for #dotace and #babis originated from other countries, such as Italy and India. Following this trend, we reframed our utility to include the "user verified" attribute, which details the reliability of the user as a consistent account. The Levenstein distance for each tweet was computed to find potential duplicates from the originating accounts. In the case of #babis, some duplicates were found, however those originated from multiple sources and it was difficult to discern an original source. None of the analyzed tweets appeared to be duplicates, however there was strong indication that other users reused the content material from older tweets concerning #babis, #dotace, #repka.

What the data tells us, is that the majority of Twitter traffic in the Czech agricultural sector is significantly oriented around rapeseed or related biofuel topics and the prime minister, Andrej Babiš (see Table 5). Also, there is a loose correlation with topics concerning government subsidies (#dotace) for rapeseed production, presumably from the European Union. With this in mind, we can refine our understanding of the Czech agricultural sector and the most relevant issues concerning rural communities and agricultural

Hashtag	Total number of accounts	Verified accounts	Tweets by verified accounts				
			Total	Quoted	Favorited	Retweeted	Replied
#agrofert	2,746	50	218	32	121	102	77
#babis	16,464	366	1,706	185	665	554	393
#bionafta	21,912	619	1,821	184	811	868	251
#capihnizdo	1,922	23	105	18	47	41	34
#dotace	21,573	601	1,719	172	823	794	358
#repka	9,962	139	384	26	268	241	69

Source: Own processing

Table 5: Total number of accounts, verified accounts and tweets.

Hashtag	Unverified accounts	Tweets by unverified accounts				
		Total	Quoted	Favorited	Retweeted	Replied
#agrofert	2,696	6,459	115	1,49	659	588
#babis	16,098	49,642	1,274	9,827	5,03	3,916
#bionafta	21,293	69,289	930	11,506	11,118	1,796
#capihnizdo	1,899	4,669	173	1,121	539	470
#dotace	20,972	35,252	432	5,199	4,326	1,711
#repka	9,823	23,452	391	4,207	3,107	991

Source: Own processing

Table 6: Total number of unverified accounts and tweets.

stakeholders. Furthermore, we did find several outliers concerning the proportion of verified users versus unverified users that utilize the hashtags for their own twitter content.

From the collected dataset we were able to find out how many accounts were validated, and how many were not. We can assume that the verified accounts should be more trustworthy than the accounts that are not verified (see Table 6).

There is the possibility that some non-verified accounts could be bots, however we did not have the means to detect this. In our case, it does mean that there was a disproportionate number 'of tweets that came from unverified users rather than verified users, indicating a behavior of anonymity in the activity of reusing tweets, e.g.: #babis or #bionafta. This is not necessarily uncommon, as most Twitter users are unverified. However, it is unusual in this case considering that the unverified tweets are specifically oriented around #bionafta, #dotace, #repka and #Babis. These topics seem to be commented on disproportionately by unverified users in contrast to verified users.

Conclusion

Our paper presents first analysis of Czech Twitter content within the agriculture context. We now know which topics are important to the Czech agricultural sector in terms of social media dispersion. However, considering we only were able to capture 240,000 tweets over a five-year period, we cannot say that Twitter is an important medium for capturing the complete range of stakeholders. Nonetheless, if even the proportion of tweets analyzed is indicative of social media activity surrounding Czech agriculture, then we can assume that the hashtags were correct as far as the general pulse of the sector is concerned. A follow up will include a semantic analysis of the tweets and exploration of other social media platforms utilized in the Czech Republic for a larger sample size. At this stage, we cannot say what number of unverified accounts are bots, as that will require new analytical techniques such as Latent Dirichlet Allocation and dynamic topic modelling. However, we are in a better position to follow up with a second stage of analysis that will include the aforementioned methods and techniques.

Acknowledgements

This work was supported by the Internal Grant Agency of the Faculty of Economics and Management, Czech University of Life Sciences Prague under Grant number 2019A0010.

Corresponding authors
John Phillip Sabou, MSc.
Department of Information Technologies, Faculty of Economics and Management
Czech University of Life Sciences in Prague
Kamýcká 129, 165 00 Prague - Suchdol, Czech Republic
E-mail: sabou@pef.czu.cz

References

[1] Bordag, S. (2008) "A comparison of co-occurrence and similarity measures as simulations of context", *International Conference on Intelligent Text Processing and Computational Linguistics*. pp 52-63. ISSN 0302-9743. DOI 10.1007/978-3-540-78135-6_5.

[2] Campoy, A. (2019) "*More than 60% of Donald Trump's Twitter followers look suspiciously fake*". [Online]. Available: https://qz.com/1422395/how-many-of-donald-trumps-twitter-followers-are-fake/ [Accessed: 28 Nov. 2019].

[3] Carré, P. and Pouzet, A. (2014) "Rapeseed market, worldwide and in Europe", *Oilseeds and fats, Crops and Lipids*, Vol. 21, No. 1. E-ISSN 2257-6614, ISSN 2272-6977. DOI 10.1051/ocl/2013054.

[4] Červenková, E., Šimek, P., Vogeltanzová, T. and Stočes, M., (2011) "Social networks as an integration tool in rural areas–agricultural enterprises of the Czech Republic", *AGRIS on-line Papers in Economics and Informatics*, Vol. 3, No. 1, pp 53-60. ISSN 1804–1930.

[5] Church, K. W. and Hanks, P. (1990) „Word association norms, mutual information, and lexicography", Proceeding of ACL ,89 Proceedings of the 27th annual meeting on Association for Computational Linguistics, pp. 76-83. DOI 10.3115/981623.981633.

[6] Dickerson, J. P., Kagan, V. and Subrahmanian, V. S. (2014) "Using sentiment to detect bots on twitter: Are humans more opinionated than bots?", *In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp 620-627. IEEE Press. E-ISBN 978-1-4799-5877-1. DOI 10.1109/ASONAM.2014.6921650.

[7] Jansen, B. J., Zhang, M., Sobel, K. and Chowdury, A. (2009) "Twitter power: Tweets as electronic word of mouth", *Journal of the American society for information science and technology*, Vol. 60, No. 11, pp 2169-2188. E-ISSN 2330-1643. DOI 10.1002/asi.21149.

[8] Kołodziejczak, A. and Kossowski, T. (2011) "Diversification of farming systems in Poland in the years 2006-2009", *Quaestiones Geographicae*, Vol. 30, No. 2, pp 49-56. ISSN 0137-477X. DOI 10.2478/v10117-011-0017-x.

[9] Lund, K. and Burgess, C. (1996) "Producing high-dimensional semantic spaces from lexical co-occurrence", *Behavior research methods, instruments, & computers*, Vol. 28, No. 2, pp 203-208. E-ISSN 1554-3528. DOI 10.3758/BF03204766.

[10] Mazurek, G., Korzyński, P. and Górska, A. (2019) „Social Media in the Marketing of Higher Education Institutions in Poland: Preliminary Empirical Studies", *Entrepreneurial Business and Economics Review*, Vol. 7, No. 1, pp 117-133. E-ISSN 2353-8821, ISSN 2353-883X. DOI 10.15678/EBER.2019.070107.

[11] Miller, G. A. and Charles, W. G. (1991) "Contextual correlates of semantic similarity", *Language and cognitive processes*, Vol. 6, No. 1, pp 1-28. E-ISSN 2327-3801, ISSN 2327-3798. DOI 10.1080/01690969108406936.

[12] Orhan, M.A. (2017) „ The Evolution of the Virtuality Phenomenon in Organizations: A Critical Literature Review", *Entrepreneurial Business and Economics Review,* Vol. 5, No. 4, pp. 171-188. E-ISSN 2353-8821, ISSN 2353-883X. DOI 10.15678/EBER.2017.050408.

[13] Ozdikis, O., Senkul, P. and Oguztuzun, H. (2012) "Semantic expansion of hashtags for enhanced event detection in Twitter", *Conference: VLDB Workshop on Online Social Systems (WOSS 2012), Istanbul.* DOI 10.1109/ASONAM.2012.14.

[14] Pehe, J. (2018) "Czech Democracy Under Pressure", *Journal of Democracy*, Vol. 29, No. 3, pp. 65-77. E-ISSN 1045-5736, ISSN 1086-3214. DOI 10.1353/jod.2018.0045.

[15] Reiff M., Surmanová K., Balcerzak A. P. and Pietrzak M. B. (2016) „Multiple Criteria Analysis of European Union Agriculture", *Journal of International Studies*, Vol. 9, No 3, pp. 62-74. E-ISSN 2071-8330, ISSN 2306-3483. DOI 10.14254/2071-8330.2016/9-3/5.

[16] Spence, D. P. and Owens, K. C. (1990) "Lexical co-occurrence and association strength", *Journal of Psycholinguistic Research*, Vol. 19, No. 5, pp. 317-330. E-ISSN 1573-6555, ISSN 0090-6905. DOI 10.1007/BF01074363.

[17] Svatoš, M. and Smutka, L. (2009) "Influence of the EU enlargement on the agrarian foreign trade development in member states", *Agricultural Economics (AGRICECON)*, Vol. 55, No. 5, pp. 233-249. ISSN 0139-570X. DOI 10.17221/34/2009-AGRICECON.

[18] Turney, P. D. (2001) "Mining the web for synonyms: PMI-IR versus LSA on TOEFL", Proceedings of the 12ve *European Conference on Machine Learning*, pp. 491-502. Springer, Berlin, Heidelberg. DOI 10.1007/3-540-44795-4_42.

[19] Vaněk, J., Šimek, P., Vogeltanzová, T., Červenková, E. and Jarolímek, J. (2010) "ICT in Agricultural Enterprises in the Czech Republic–Exploration 2010", *Agris on-line Papers in Economics and Informatics*, Vol. 2, No. 3, pp 69-75. ISSN 1804–1930

[20] Wald, R. and Khoshgoftaar, T. M., Napolitano, A. and Sumner, C. (2013) "Predicting susceptibility to social bots on twitter", *14th International Conference on Information Reuse & Integration (IRI)*, IEEE Computer Society, pp. 6-13.

[21] Wątróbski, J., Jankowski, J. and Ziemba, P. (2016) „Multistage Performance Modelling in Digital Marketing Management", *Economics and Sociology*, Vol. 9, No. 2, pp. 101-125. ISSN 2071-789X. DOI 10.14254/2071-789X.2016/9-2/7.

[22] Xu, Z., Ru, L., Xiang, L. and Yang, Q. (2011) "Discovering user interest on twitter with a modified author-topic model", *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology,* , Vol. 1, pp. 422-429. DOI 10.1109/WI-IAT.2011.47.