Big Data and Sentiment Analysis using KNIME: Online Reviews vs. Social Media

Ana Mihanović, Hrvoje Gabelica, Živko Krstić
Poslovna inteligencija d.o.o., Zagreb, Croatia
{ana.mihanovic, hrvoje.gabelica, zivko.krstic}@inteligencija.com

Abstract - Text analytics and sentiment analysis can help an organization derive potentially valuable business insights from text-based content such as word documents, email and postings on social media streams like Facebook, Twitter and LinkedIn. The system described here analyses opinions about various gadgets collected from two different sources and in two different forms; online reviews and Twitter posts (tweets). Sentiment analysis can be applied to online reviews in easier and more detailed way than to the tweets. Namely, online reviews are written in clear and grammatically more accurate form, while in tweets, internet slang, sarcasm and allegory are often used. System described here explains methods of data collection, sentiment analysis process for online reviews and tweets using KNIME, gives an overview of differences and analysis possibilities in sentiment analysis for both data sources.

I. INTRODUCTION

Text mining or sentiment analysis [1] is analysis of data contained in a natural language text, which deals with the computation of opinion, sentiment and subjectivity in text. Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information from the text documents. Basic task of sentiment analysis is to determine the polarity of a given texts.

Tasks of dictionary making and sentiment analysis process are done by the means of KNIME [2], which is a user-friendly graphical workbench capable of entire analysis process. KNIME uses six different steps to process texts: reading and parsing documents, named entity recognition, filtering and manipulation, word counting and keyword extraction, transformation and visualization. Following workflows and tasks are developed and executed using KNIME:

- Retrieving data from database
- Dictionary development and implementation
- Review scoring

II. DATA COLLECTION

In gathering online reviews and tweets about gadgets, focus was set on few gadget manufacturers such as Apple, Samsung, Nokia, Nexus, LG and on the webpages that contain online reviews about gadgets. Total number of collected online reviews and tweets which were collected during few hours for the purposes of this paper was:

- 812176 tweets
- 419624 online reviews

The system presented here handles crawling, extracting gadget reviews and storing them for analysis. Collected unstructured text is prepared for text mining and sentiment analysis. System presented gives the analysis results for every single review or tweet.

Online reviews were crawled from webpages using Apache Nutch [3] crawler, highly extensible and scalable open source web crawler which traverses the web site starting from a given set of URLs and follows the links matching a given pattern to a certain depth. Tweets were collected with in-house developed Java package used for streaming posts from Twitter. Both online reviews and tweets were collected within few hours. Since tweets are much shorter than online reviews and collecting these posts takes less time than crawling online reviews, the number of tweets is much bigger than the number of crawled online reviews.

Crawled online reviews and tweets and stored into HBase tables on Apache Hadoop [4] server. Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers and it is popular for developing large-scale data-intensive applications. Hadoop is used as a storage environment, and HBase table, as a component of Hadoop, is used to store information about gadgets and gadgets reviews. HBase [5] is non-relational, distributed database written in Java which efficiently holds unstructured data in large tables and can be concurrently and randomly accessed. Hadoop and HBase were used because of their main functionalities, which are storage of large quantities of unstructured, textual data and their possibility of reading and writing data in the real time.

III. DATASET DESCRIPTION

Data collected into HBase tables slightly differs for online reviews and tweets.

Online reviews are described with following attributes and structured in the following way:

- Key - unique key indicator for each gadget review created before review was stored into database
- PID (Product group ID) – name of the gadget, such as Samsung Galaxy s4 or iPhone 5
- Review Date – date when online review was stored into HBase table
- Review text – text of the review about specific gadget
- Lang – language of the review text

Tweets are described with following attributes and structured in the following way:

- Key - unique key indicator for each gadget review created before tweet was stored into database
- userScreenName – Username of Twitter user
- CreationDate – date when tweet was stored into HBase table
- text – tweet text about specific gadget
- keyword – name of the gadget, such as Samsung Galaxy s4 or iPhone 5, analogous to PID field in online reviews table
- Lang – language of the tweet

Key and PID value (keyword vale for tweets) allowed aggregated analysis on the level of each review or tweet, or on the level of the specific gadget. Date values allowed aggregated analysis for the defined period of time.

Described data is loaded into KNIME with HBase Reader node and processed. In this phase, only online reviews and tweets in English language were collected. Language of the text is set to English and all texts that have different language values are filtered out, because English dictionary applied on reviews and posts written in other languages would not give results.

IV. DICTIONARY BUILDING

A. Online reviews

Online reviews are analyzed on the category level. Seven different categories are introduced for the gadget reviews:

- Accessibility/usage – ease of usage, availability of applications and software
- Value for money – is the price of the product reasonable or not, is the price justified by quality
- Content/composition – materials that product is made of, all hardware components, accessories (handset, bluetooth handset, keypad, etc.)
- Quality – quality of materials product is made of, quality of hardware components, manufacturer quality
- User experience – experience of the user when buying and using product (issues, would recommend, wouldn’t recommend, would buy again, etc.)
- Look/appearance – physical appearance of the product (color, design, size, weight, etc.)
- Service/Support – seller’s attitude and kindness, service center help support, warranty of the product

Service/support category is the category with the least number of the recognized phrases and for that reasons it is not included in chart visualizations. Redefiniton of this category is being considered.

Dictionary building for detailed sentiment analysis implies making an initial list of adjectives and nouns which are normally used when describing a specific product. Initial list is collected by analyzing existing opinions and reviews and manually extracting the words which appear the most in those reviews and opinions or those words that seem to be most important. Grade scope is defined from 0 to 5 (0 meaning extremely dissatisfied, and 5 meaning extremely satisfied). Null means no rating.

Since terms and phrases extracted from reviews need to have category and grade, nouns are holders of categories, and adjectives are holder of grades. That means that every noun on the list had to be categorized into one of seven categories, and every adjective on the list has to be graded with one grade. Once the initial list of nouns and adjectives has been collected, nouns and adjectives are used in regular expressions which bind them together and recognize complete phrases in reviews consisted of all combinations of adjectives and nouns that are in the list. Regular expressions are composed for the most used grammatical forms in English language. List of regular expressions used for dictionary development:

- Adjective + noun and negations („not“ + adjective + noun)

This form is used to detect most simple phrases of English language, such as „amazing battery“, „bad screen“ or „great apps“. For the negation detection, regular expressions are written in the form „not“ + adjective + noun.

- Noun + „to be“ + adjective and negations (noun + „not to be“ + adjective)

With these forms of regular expressions, more complicated phrases are detected, such as „battery doesn’t last“, „screen is great“, or „camera is good“. To make phrases detection more accurate and successful, all forms of verb „to be“ must be included, including grammatical mistakes which are people likely to make.

- Noun + intensifier/downtoner + adjective and negations (noun + „not to be“ + intensifier/downtoner + adjective)

Intensifies or amplifiers are words in English that increase the effect of the verb and include such words as „completely“, „totally“, „undoubtedly“, „absolutely“, etc. Downtoners are words that decrease the effect of the verb and include such words as „kind of“, „not so much“, „sort of“, and so on.

Another important feature is to correctly grade phrases that contain negations and/or modifiers because these type of words change total grade and meaning of a phrase. If phrase contains some of defined words, grade is calculated using different “if...else if” rules. These exceptions are handled as follows:

- If the phrase contains negation, the grade is reversed and replaced by its module (e.g. „camera is great“=4, „camera isn’t great“=1)
- If the phrase contains intensifier, the grade is increased by one (e.g. „good processor “=3, „really good processor“=4). If the phrase contains negations along with intensifier, the grade module is calculated (e.g. „not really good processor“ =1)
- If the phrase contains downtoner, the grade is decreased by one (e.g. „display is good“=3, „display is merely good“=2). Negations used along with downtoners are rare and therefore not analyzed

B. *Twitter posts*

Dictionary built for sentiment analysis of tweets contains of key words graded only as positive or negative. Scoring or sentiment analysis of tweets is done on the positive-negative level, because tweets don’t contain clear phrases such as those that can be found in online reviews. Therefore, tweets are impossible to categorize into numerous categories as online reviews, and they need to be analyzed on the word level, giving each word positive or negative polarity.

Dictionary built for sentiment analysis of social media posts consist of most used words in those posts. Because of the different structure when comparing online reviews and tweets, dictionary needs to include jargon words, internet slang, smiley icons, abbreviations, hashtags and similar. These words and symbols are of great importance, since tweets are usually not rich with useful phrases and terms. Example of such words included into dictionary are: #loveit, #horror, :-(, :-*, OMG, JK and similar.

For this task, publicly available MPQA subjectivity lexicon was used as a starting point for recognizing contextual polarity [6], which was expanded with Twitter specific words mentioned above. Existing dictionary containing of approximately 8000 words is expanded to fit the needs for gadget analysis in a way that initial portion of tweets are collected, which are separated into single words with Bag of Words processing. Unnecessary words such as symbols or web URLs are filtered out, and all useful, social media specific words are graded and added to the dictionary.

V. DATA SCORING

Sentiment analysis of online reviews and tweets differs greatly. Online reviews are easier to objectively analyze because of clearer written form and bigger amount of meaningful sentiments, but they have to be analyzed on a more detailed level than tweets.

Tweets often contain internet slang, sarcasm and allegory which are often used. It could be said that grading online reviews is easier, but the dictionary is more complicate to make, and that grading tweets is harder, but the dictionary is easier to make.

Online reviews are analyzed on the phrase and category level, giving phrase a grade for one of seven categories. Tweets are analyzed on the word level giving a positive or negative grade for each term. Text analytics of online reviews is accomplished simply with phrases counters and mean calculations, while analytics of tweets is frequency-driven.

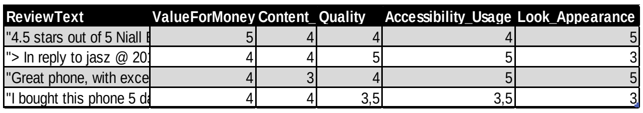
Big quantity of reviews or posts are loaded into KNIME to be graded. Two separate wokflows are built, one for grading onine reviews based on a grade and category, and other one for positive-negative grading.

A. *Online Reviews Grading*

First input of sentiment analysis workflow developed in KNIME are online reviews read from the database, and second, parallel input is the dictionary made of phrases recognized with regular expressions. Phrases from the dictionary are recognized from the dictionary file and tagged in Review Texts using Dictionary Tagger node.

Online reviews are processed and graded with the term presence method [7], rather than term frequency method. Term presence method gives a binary value which simply indicates does the term or phrase occur in the text (value 1) or not (value 0). Every term or phrase has also a grade and category joined and binary values of terms are summed on the level of each review, giving term counters and grade sum for each review.

Sum of grades for every category is divided with counter number for every category, giving final grade for each online review and for each category. The results were written back to the Hadoop database (Fig. 1), and as a result of Hadoop aggregation, the average grade for every category on the level of single gadget is calculated.



ReviewText	ValueForMoney	Content_	Quality	Accessibility_Usage	Look_Appearance
"4.5 stars out of 5 Niall E	5	4	4	4	5
"> In reply to jasz @ 20	4	4	5	5	3
"Great phone, with exce	4	3	4	5	5
"I bought this phone 5 da	4	4	3,5	3,5	3

Figure 1. Example results of analyzed online reviews

B. *Twitter Posts Grading*

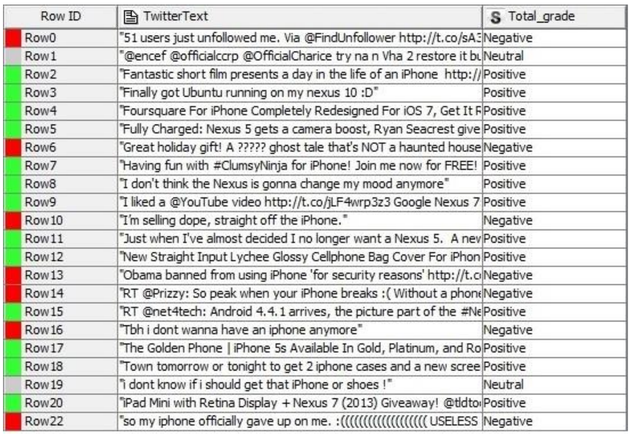
Tweets are analyzed only on the positive-negative level. The result can be one of the three; positive, negative and neutral. Tweets for gadgets are not analyzed on the category level because these kind of text documents are often much shorter and not so descriptive as online reviews. Tweets are processed in similar way, but this approach uses more preprocessing steps and uses frequency-driven approach. In the preprocessing steps, various filters are applied on the text documents, such as stopwords filter, n-chars filter and punctuation erasure.

TF*IDF (Term Frequency*Inverse Document Frequency) [8] method assigns non-binary weights related on a number od occurences of a word. Weighting exploits counts from a background corpus, which is a large collection of documents; the background corpus serves as indication of how often a word may be expected to appear in an arbitrary text. TF*IDF calculation determines how relevant a given word is in a particular document [9].

Besides term frequency $f_{w,d}$ which equals the number of times word w appears in document d, size of the corpus D is also needed. Given a document collection D, a word w and an individual document $d \epsilon D$, TF*IDF value can be calculated [9]:

$$TF * IDF_w = f_{w,d} * \log \frac{D}{f_{w,d}} \tag{1}$$

Total score for each word is given by multiplying TF*IDF value with attitude of a term. Attitude can have one of three values depending on the word polarity; -1 for word with negative polarity, +1 for word with positive polarity and 0 for neutral words. Final weigths, which now represent attitude of each document , are grouped on the level of document (tweet) and binned into three bins to give one of three final results for each tweet; positive, negative or neutral (Fig. 2).



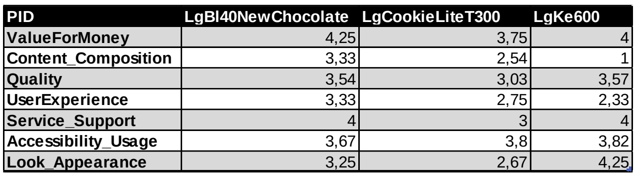
Row ID	TwitterText	Total_grade
🟥 Row0	"51 users just unfollowed me. Via @FindUnfollower http://t.co/sA3	Negative
⬜ Row1	"@encef @officialccrp @OfficialCharice try na n Vha 2 restore it bu	Neutral
🟩 Row2	"Fantastic short film presents a day in the life of an iPhone http://	Positive
🟩 Row3	"Finally got Ubuntu running on my nexus 10 :D"	Positive
🟩 Row4	"Foursquare For iPhone Completely Redesigned For iOS 7, Get It F	Positive
🟩 Row5	"Fully Charged: Nexus 5 gets a camera boost, Ryan Seacrest give	Positive
🟥 Row6	"Great holiday gift! A ????? ghost tale that's NOT a haunted house	Negative
🟩 Row7	"Having fun with #ClumsyNinja for iPhone! Join me now for FREE!	Positive
🟩 Row8	"I don't think the Nexus is gonna change my mood anymore"	Positive
🟩 Row9	"I liked a @YouTube video http://t.co/jLF4wrp3z3 Google Nexus 7	Positive
🟥 Row10	"I'm selling dope, straight off the iPhone."	Negative
🟩 Row11	"Just when I've almost decided I no longer want a Nexus 5. A nev	Positive
🟩 Row12	"New Straight Input Lychee Glossy Cellphone Bag Cover For iPhon	Positive
🟥 Row13	"Obama banned from using iPhone 'for security reasons' http://t.c	Negative
🟥 Row14	"RT @Prizzy: So peak when your iPhone breaks :(Without a phone	Negative
🟩 Row15	"RT @net4tech: Android 4.4.1 arrives, the picture part of the #Ne	Positive
🟥 Row16	"Tbh i dont wanna have an iphone anymore"	Negative
🟩 Row17	"The Golden Phone | iPhone 5s Available In Gold, Platinum, and Ro	Positive
🟩 Row18	"Town tomorrow or tonight to get 2 iphone cases and a new scree	Positive
⬜ Row19	"i dont know if i should get that iPhone or shoes !"	Neutral
🟩 Row20	"iPad Mini with Retina Display + Nexus 7 (2013) Giveaway! @tldto	Positive
🟥 Row22	"so my iphone officially gave up on me. :((((((((((((((((((((USELESS	Negative

Figure 2. Results of analyzed tweets

VI. RESULTS

A. *Online reviews results*

Online reviews analysis counts all terms in one reviews, and gives mean grade on the review level, for every category. Results can be aggregated on the level of gadget (Fig. 3) or specific model and monitored over time time to analyze trendings.



PID	LgBl40NewChocolate	LgCookieLiteT300	LgKe600
ValueForMoney	4,25	3,75	4
Content_Composition	3,33	2,54	1
Quality	3,54	3,03	3,57
UserExperience	3,33	2,75	2,33
Service_Support	4	3	4
Accessibility_Usage	3,67	3,8	3,82
Look_Appearance	3,25	2,67	4,25

Figure 3. Analyzed online reviews results aggregated on the level of LG products

Simple tag cloud (Fig. 4) gives the overview of the most used phrases in online reviews:

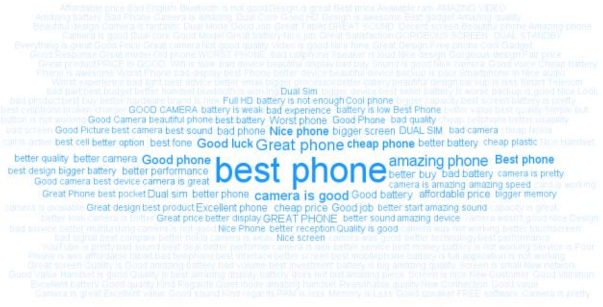


Figure 4. Tag cloud for online reviews

B. *Twitter posts results*

Tweets analysis results are given in the form of positive, negative and neutral. Neutral grade can be avoided, which is easily accomplished by changing grade bins and removing a bin for neutral grade. Grade can be calculated on the level of specific post and then aggregated for all posts of one gadget product (Fig. 5). Sample of 40000 tweets were used for iPhone product analysis.

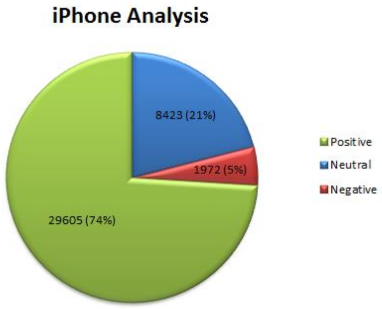


Figure 5. Analyzed tweets results aggregated on the level of gadget brand

Simple tag cloud (Fig. 6) gives the overview of the most used words in tweets:

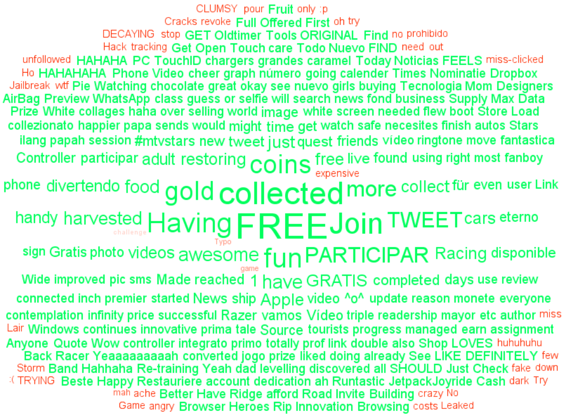
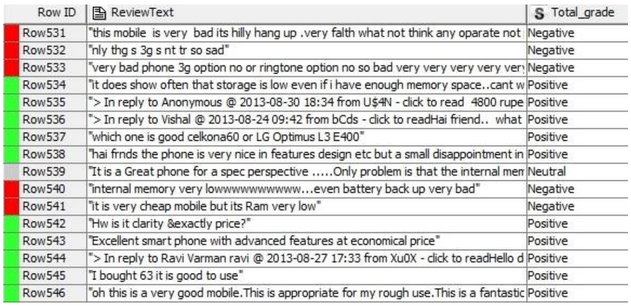


Figure 6. Tag cloud for tweets

C. *Online reviews graded with Twitter dictionary*

When online reviews are graded with the dictionary made for analysis of tweets (Fig. 7), results are fairly good because each word is separately processed and graded. Dictionary developed for sentiment analysis of tweets is usable for the sentiment analysis of online reviews.

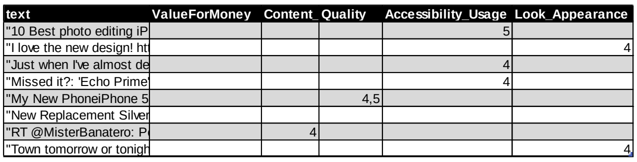


Row ID	ReviewText	Total_grade
🟥 Row531	"this mobile is very bad its hilly hang up .very falth what not think any oparate not	Negative
🟥 Row532	"nly thg s 3g s nt tr so sad"	Negative
🟥 Row533	"very bad phone 3g option no or ringtone option no so bad very very very very ver	Negative
🟩 Row534	"it does show often that storage is low even if i have enough memory space..cant w	Positive
🟩 Row535	"> In reply to Anonymous @ 2013-08-30 18:34 from U$4N - click to read 4800 rupe	Positive
🟩 Row536	"> In reply to Vishal @ 2013-08-24 09:42 from bCds - click to readHai friend.. what	Positive
🟩 Row537	"which one is good celkona60 or LG Optimus L3 E400"	Positive
🟩 Row538	"hai frnds the phone is very nice in features design etc but a small disappointment in	Positive
⬜ Row539	"It is a Great phone for a spec perspective .....Only problem is that the internal mem	Neutral
🟥 Row540	"internal memory very lowwwwwwwwwww....even battery back up very bad"	Negative
🟥 Row541	"it is very cheap mobile but its Ram very low"	Negative
🟩 Row542	"Hw is it clarity &exactly price?"	Positive
🟩 Row543	"Excellent smart phone with advanced features at economical price"	Positive
🟩 Row544	"> In reply to Ravi Varman ravi @ 2013-08-27 17:33 from Xu0X - click to readHello d	Positive
🟩 Row545	"I bought 63 it is good to use"	Positive
🟩 Row546	"oh this is a very good mobile.This is appropriate for my rough use.This is a fantastic	Positive

Figure 7. Results of online reviews graded with the Twitter dictionary

D. *Twitter posts graded with online reviews dictionary*

Tweets graded with dictionary developed for sentiment analysis of online reviews (Fig. 8) are not succesfuly analyzed because of absence of the more complicated phrases that tweets lack. Therefore, tweetscan't be graded on the detailed level, and positive-negative analysis or analysis with simple grade from 1 to 5 is most suitable for sentiment analysis of social media.



text	ValueForMoney	Content_	Quality	Accessibility_Usage	Look_Appearance
"10 Best photo editing iP				5	
"I love the new design! htt					4
"Just when I've almost de				4	
"Missed it?: 'Echo Prime'				4	
"My New PhoneiPhone 5			4,5		
"New Replacement Silver					
"RT @MisterBanatero: P		4			
"Town tomorrow or tonigh					4

Figure 8. Results of tweets analysis graded with the online reviews dictionary

VII. CONCLUSION

Sentiment analysis of online reviews is less complicated process and gives more detailed results, but developing a dictionary for these kind of text documents takes more time and resources than for social media.

On the other side, sentiment analysis of social media posts is more complicated process, but dictionary development is less complicated task. Also, dictionary developed for sentiment analysis of social media, on a single word level, can be successfuly applied on sentiment analysis of online reviews, but vice versa situation does not give results. Social media posts are hard to analyze on the phrase or sentence level because of theirs unique structure and grammar.

[1] B. Pang, L. Lee, “Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval”, vol. 2, No. 1–2, 2008, pp 1-135

[2] http://knime.org/

[3] http://nutch.apache.org/

[4] http://hadoop.apache.org/

[5] http://hbase.apache.org/

[6] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, University of Pittsburgh, Pennsylvania, 2005

[7] J. Kim, J. Li, J. Lee, “Discovering the Disciriminative Views: Measuring Term Weights for Sentiment Analysis”, Pohang University of Science and Technology, Republic of Korea, 2009, pp 253-261

[8] C. C. Aggarwal, C. Zhai, Mining Text Data, Springer Science+Business Media, New York, 2012, pp 48-52

[9] J. Ramos, “Using TF-IDF to Determine Word Relevance in Document Queries”, Rutgers University, New Jersey, 2012