International Workshop on Curbing Collusive Cyber-gossips in Social Networks (C3-2014), August 17, 2014
Proc. IEEE/ACM Int. Conf. Advances in Social Network Analysis and Mining, ASONAM, Beijing, China.

Simpler is Better? Lexicon-based Ensemble Sentiment Classification Beats Supervised Methods

Lukasz Augustyniak, Tomasz Kajdanowicz, Piotr Szymański, Włodzimierz Tuligłowicz,
Przemyslaw Kazienko, Reda Alhajj, Boleslaw Szymanski
Wroclaw University of Technology, Institute of Informatics, Wroclaw, Poland
Email: {lukasz.augustyniak, tomasz.kajdanowicz, piotr.szymanski, wlodzimierz.tuliglowicz, przemyslaw.kazienko}
@pwr.edu.pl, alhajj@ucalgary.ca, szymab@rpi.edu

***Abstract*—It has been shown in this paper that simplistic Bag of Words (BoW) lexicon methods for sentiment polarity assignment with ensemble classifiers are much faster than a supervised approach to sentiment classification while yielding similar accuracy. BoW methods also proved to be efficient and fast across all examined datasets. Moreover, a new approach to lexicon extraction that can be successfully used for sentiment polarity assignment is presented in the paper. It has been shown that accuracy obtained from such lexicons outperforms other lexicon based approaches.**

I. Introduction

Ever since the beginning of trade companies used opinions and reviews as a method of building trust in new clients or gathering feedback from old clients. Was it ancient Rome, mid XIX-century French middle-class entrepreneurs or the modern e-commerce, user feedback had been an invaluable element of every commercial success. Yet it was the rapid development of internet-based markets that allowed massive opinion collection. Such data opened the doors to automatic recognition, classification and even mining opinions and their sentiment, i.e. providing computer generated information about what are the prevalent opinions about a product the company is selling or manufacturing.

With the stream being so large digesting so much more detailed information, while still important, become secondary. The primary information expected is - what is the feedback to our action? How many people said something positive? How many were strongly against? Was someone indifferent? Thus we come close to the basic model of sentiment classification - deciding whether a given text is positive, negative or neutral.

Many different attempts have been made at solving the sentiment classification problem. From selecting a very few positive/negative words for basic classification, through large sentiment lexicons, to word-relationship based approaches such as sentinet. Yet while the stream of user generated content flows stronger than ever, it is built from the same finite set of building blocks. Electronic opinions usually consist of a quantitative indication of how satisfied a user has been (such as grades, stars, likes/dislikes, sometimes a named emotion indication) and textual description of the product experience.

Both of those user-experience feedback methods hold up to a standard of comparability. Stars/grades/likes/dislikes can be translated into a standard scale on the $[-1, 1]$ interval and the textual resources are still composed of a similar, common subset of words in a given language. This enables searching for correlations of certain words being used more often than others in reviews of a given sentiment. This approach yielded the search for a good indicator of possible correlations. The original naive word-to-sentiment frequency proved not to be resilient enough to changes in number of reviews per sentiment or to switching from one topic to another.

Lexicon building is thus a hard task in terms of avoiding numerical and statistical artifacts occurring due to certain characteristics of data sampling methods. A way of overcoming this problem was to use supervised learning methods, which would learn of provided data sets to differentiate the input space well enough to infer a text's sentiment. Such methods often secure results without the need to understand the nature of input data, yet remain more intractable for ever growing data sets. Supervised methods advantage lies in the fact that they are provided as ready-to-use functions in many machine learning libraries, while constructing lexicon-based methods requires more work to setup.

We have combined the best of both worlds - a mixture of a frequency-related measure used to build lexicons and an ensemble method to infer from lexicon based results. Such a combination provided efficiency on par with supervised learning while providing a major speed improvement.

To provide an insight into the state of art we describe the historical attempts and related methods in the Section II-A. We provide an introduction into ensemble classifier methods in subsection II-C. We then proceed to describe our lexicon-based classifier ensemble approach to the problem in Section III. In Section IV we describe utilized data set, present experimental results and, finally, discuss them in Section V.

II. Related Work

A. *Sentiment Analysis*

The area of sentiment analysis (also called sentiment extraction, sentiment mining, opinion mining) has been an object of interest of many authors in recent years. Liu and Zhang [1] described three main approaches to sentiment analysis: lexicon-based method, classification with supervised and unsupervised learning methods.

Hatzivassiloglu and McKeown [2] introduced an algorithm of building a sentiment lexicon based on word corpora. The method uses the fact that adjectives of the same semantic orientation are more likely conjoined by the word ”and” or ”either ... or ...” then the word ”but” or ”neither ... nor ...”. They used also morphological dependencies between adjectives (e.g. words ”thoughtful” and ”thoughtless” have opposite sentiment orientation). Their method is accurate, but designed for isolated adjectives. In general, lexicon-based methods are probably the simplest, but have many shortcomings, as described in [3].

Second group of methods is using a classifier such as Naive Bayes, SVM, Decision Tree or other. Pang et al. [4] researched which features of text used in supervised learning classification help getting better results and also what kind of preprocessing of text helps to improve the final accuracy.

Turney [5] and Liu et al. [1] shown an unsupervised learning method of predicting sentiment similar to lexicon-based. The prediction in his method is dependent on average sentiment orientation of each opinion word in the text, but the sentiment of each word is not taken from any conventional sentiment lexicon, but from number of results found by an AltaVista Advanced search engine from specially designed queries. Turney achieved an average accuracy of 74% of predicting the sentiment of reviews from different topics, ranging from 66% for movie reviews to 84% for automobile reviews.

B. Sentiment Analysis of User-Generated Content

Sentiment Analysis is a hard problem on its own, but when it comes to analyzing text from social media things get even harder. Typos, intentional misspellings, emoticons, jargon are just few additional obstacles in the task [4]. Another, more complex obstacle is sarcasm, that is ”saying the opposite of the truth, or the opposite of their true feelings in order to be funny or to make a point”[1]. Maynard [6] had investigated an impact of sarcasm on sentiment analysis of tweets. She used hashtags to determine if the tweet is sarcastic or not. One of the conclusions of her work was that even with such a simplification in detecting sarcasm as hashtags it is still hard to predict the overall sentiment orientation of a tweet.

Preotiuc-Pietro et al. [7] have shown that in case of analyzing sentiment in text from social media it is possible that accuracy of sentiment analysis to drop significantly when using standard preprocessing tools such as part-of-speech taggers or named entity recognition systems. That is due to the fact that these kind of texts are often noisy and conversational [7].

C. Ensemble classifiers

Ensemble learning technique combines multiple weak learners in an attempt to produce one strong learner.

Whitehead [8] describes ensemble learning as a technique increasing machine learning accuracy with a trade-off of increasing computation time so they are best suited in those domains where computational complexity is relatively unimportant compared to the best possible accuracy.

One of the first ensemble learning techniques was bootstrap aggregating (shorter: bagging). As described in Breiman [9] bagging technique involves generating (by using bootstrap replicates of the training set) multiple versions of a predictor and using them to form one aggregated predictor. Tests on real and simulated data sets show that bagging can give substantial gains in accuracy.

Another ensemble method is called boosting. Schapire [10] presented its basic idea as consisting of three steps: (1) performing an iterative search to locate the regions/examples that are more difficult to predict, (2) rewarding accurate predictions on those regions in each iteration, (3) combining the rules from each iteration. He also presented his version of boosting algorithm, called AdaBoost (short for adaptive boosting), which solved many of the practical difficulties of its predecessor.

There are many other ensemble algorithms. The variety of these algorithms is caused by the difference in answers to the three basic questions presented in Polikar [11]: (1) How are subsets of the training data chosen for each individual learner? (2) What types of learners are used to form the ensemble? (3) How are classifications made by the different individual learners combined to form the final prediction?

III. Proposed method

In previous paper [12] it has been noticed that for sentiment analysis the lexicon-based approach is much faster than the supervised learning methods. However, lexicons used individually are not good enough to assign sentiment. Hence, the idea is to improve the efficiency of lexicon-based method by use of several lexicons, and by assembling classification provided by these simple and fast methods. In this section the new sentiment lexicon extraction and ensemble method for sentiment classification is presented.

A. A New Approach to Lexicon Generation

We propose a bag-of-word/lexicon-based ensemble method where lexicon-based BoW weak learners are used to provide input for a stronger decision tree based learner.

The first step of the method is preparing the lexicons. We have decided to employ a variety of lexicons starting from simplest word lists, through word-frequency based generated word lists to established lexicons.

We begin with a very basic 2-word list consisting of strong sentiment words - good/bad, and we call the method based on it - SM - simplest method. For reference purposes we used lexicons, which we called SL/SL+ - simple lists - which are based on a recently presented lexicon [13] based on IMDB review data [14]. Another lexicon comes from our not yet thoroughly discussed intuitions that the tense of verbs is correlated with opinion’s sentiment. This lexicon consists of a short list of basic English verbs conjugated in different tenses - we call it PF. In general, we have observed that positive opinions are more frequently expressed with present and future tenses, whereas negative with past tenses. The PF+ lexicon is a sum of PF and SL+. By Bing Liu lexicon we denote the Bing Liu Opinion Lexicon[1].

We have supplemented those lexicons with additional word lists, which we call WL, 5MF and 25MF respectively explained below. We have assumed that the input review sets form a probability space where the sample space $\mathcal{R}$ consists of reviews represented as pairs $(score, text)$ where $text$ is represented as a set of words occurring in the review text and

[1] http://www.bbc.co.uk/worldservice/learningenglish/radio/specials/1210_how_to_converse/page13.shtml

$score$ is the normalized $[-1, 1]$ sentiment of the review. The event space is $2^{\mathcal{R}}$ and the probability function is the standard discrete probability mass function.

These lexicons were created by applying the following measure of word's w positivity/negativity, which we call frequentiment $fqmt$, defined as follows:

$$fqmt(w) = \sum_{s \in scores} s \times \frac{P(\text{review has score } s | \text{review has word } w)}{P(\text{review has score } s)} \tag{1}$$

where $score$ is a countable subset of $[-1, 1]$

The lexicon WL is a list of most 25 positive (highest $fqmt$) and 25 most negative (lowest $fqmt$) words in the input set obtained by merging all corpora.

The lexicons 5MF and 25MF were calculated for each corpus separately, selecting respectively 5 and 25 most positive (highest $fqmt$) and most negative (lowest $fqmt$) words per corpus.

B. Lexicon Based Ensemble Classification

All of the described lexicons were used for assigning sentiment polarity based on Bag-of-Words (BoW) model. This model takes individual words w in a document as features, assuming their conditional independence. For each word w in review text $t = \{w_1, w_2, ..., w_k\}$ this unigram model predicts document sentiment based on occurrence of words from lexicons. Each word w that occurs in a lexicon l has a numeric value, i.e., "1" for positive word and "-1" for negative. For words outside lexicon value "0" is assigned.

Let:

$$pos(l, t) = \# \text{ of positive words from } l \text{ that occur in t} \tag{2}$$

$$neg(l, t) = \# \text{ of negative words from } l \text{ that occur in t} \tag{3}$$

$$sum(l, t) = pos(l, t) - neg(l, t) \tag{4}$$

Then the sentiment $s_l(t)$ of a review with text t under a lexicon l is assigned using the following formula:

$$s_l(t) = \begin{cases} 1 & if\ sum(l, t) > 0 \\ -1 & if\ sum(l, t) < 0 \\ 0 & otherwise \end{cases} \tag{5}$$

Thus we obtain a sentiment polarity matrix for an input set of documents $\mathcal{D} = \{d_1, \ldots, d_n\}$ and the defined set of lexicons $\mathcal{L} = \{l_1, \ldots, l_m\}$:

$$\mathcal{S}(\mathcal{L}, \mathcal{D}) = \begin{pmatrix} s_{l_1}(d_1) & s_{l_1}(d_2) & \cdots & s_{l_1}(d_n) \\ s_{l_2}(d_1) & s_{l_2}(d_2) & \cdots & s_{l_2}(d_n) \\ \vdots & \vdots & \ddots & \vdots \\ s_{l_m}(d_1) & s_{l_m}(d_2) & \cdots & s_{l_m}(d_n) \end{pmatrix} \tag{6}$$

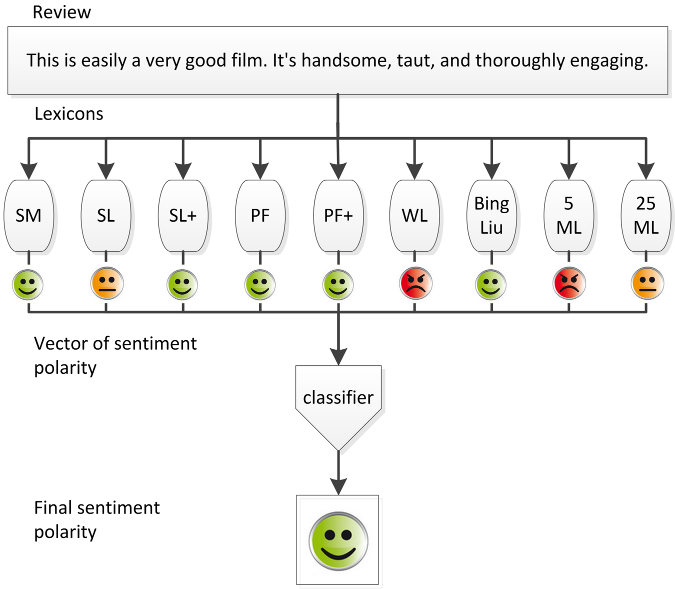
Review
This is easily a very good film. It's handsome, taut, and thoroughly engaging.
Lexicons
SM
SL
SL+
PF
PF+
WL
Bing Liu
5 ML
25 ML
Vector of sentiment polarity
classifier
Final sentiment polarity

Fig. 1. The concept of lexicon based ensemble classification of sentiment polarity for an individual review.

In the final step of our method we train a strong classifier, such as the C4.5 decision tree method, on the sentiment polarity matrix $\mathcal{S}$ with respective document texts and use such trained classifier to predict the sentiment of new documents. The diagram of the whole ensemble classification is visually presented in Figure 1.

IV. Experiments

A. Data set

In order to perform the experimental validation of the proposed methods, we are using a subset of the Amazon Reviews data set published by SNAP [15]. We have chosen these reviews from following domains:

- Automotive (188,728 reviews)
- Book (12,886,488 reviews)
- Electronics (1,241,778 reviews)
- Health (428,781 reviews)
- Movies (7,850,072 reviews)

In each of the reviews, Amazon users were stating their opinion on particular product using textual form and a discrete grade on a 1-5 scale, where 1 is the worst score and 5 is the best, and 3 can be a treated as neutral score.

The review data set has been cleaned up from its raw form. All the HTML tags and entities were stripped or converted to textual representations using the HTML parser in python library BeautifulSoup4 . Next the unicode review texts were decoded to ASCII using the unidecode python library. All punctuations from the review text were removed. Words shorter than 3 characters were discarded.

From each of the dataset we have selected reviews that were longer that 100 characters. In order to perform effective

TABLE I. STAR RATING MAPPING TO SENTIMENT CLASSES.

Star scores	Sentiment class
★	Negative
★★	Negative
★★★	Neutral
★★★★	Positive
★★★★★	Positive

computation we have limited the size of data set in each domain to the random 1600 reviews starred labelled with 1, 2, 4 or 5 stars each and 3200 labelled with 3 stars. We have thus obtained a balanced set of 3 200 positive, negative and neutral reviews each, summing to a total 9 600 reviews.

In order to check the accuracy of classified sentiment, the ground truth sentiment was extracted from ratings expressed with stars. Ratings were mapped to text classes ”positive”, ”neutral” and ”negative”, using 1 and 2 stars, 3 stars, 4 and 5 stars respectively, see Table I.

B. Experimental Scenarios

Lexicons presented in Table II have been employed in the experiments. The outputs of proposed lexicons have been used as inputs for the ensemble classifier. The experiments were carried out using Python language for text processing. Text had been cleaned before sentiment polarity assignment. Then extraction of datasets from Amazon big data sets had been done. Next, appropriate flows were created, i.e., mapping rating star scores to sentiment polarity, calculating sentiments polarity based on lexicons, concatenating outputs for each lexicon, passing these outputs to ensemble models. Decision Tree classifier with 10-fold cross validation was used for ensemble learning step. In case of this approach, 9 of 10 randomly established parts of dataset were used for lexicon creation and the 1/10 rest for evaluation.

As a baseline, basic supervised method was implemented. It involves creating feature vectors in which each attribute is derived from the whole preprocessed corpora. The attributes inform if the given word was present in the given review or not. Naturally values of each attribute can be either 0 or 1. This basic supervised method was then compared with lexicon-based ensemble method. The classification was performed using C4.5 Decision Tree. In order to evaluate this approach 10-fold cross validation was used. The results for both approaches are presented in Tables III and IV. Each approach was examined in terms of mean accuracy, precision, recall and F-measure from all cross-validation folds. Additionally, execution time was measured for each approach. The starting point was at the beginning of corpus loading and end point was at the completion of the last fold of cross-validation.

C. Results

1) Classification Accuracy Comparison of Lexicon-based Ensemble vs. Supervised Learning: As it can be observed in Table III the results measured with accuracy, precision, recall and f-measure for these two distinct approaches do not indicate superiority of any of the methods. The same is more visible in Figures 3-7, where for five distinct domains Lexicon-based Ensemble and Supervised Learning technique obtained similar results. In order to check the exact differences measured with statistical importance, the Nemenyi post-hoc test was performed and presented in Figure 2. It can be concluded that across all examined domains there is no statistical difference between Lexicon-based Ensemble and Supervised Learning approaches.

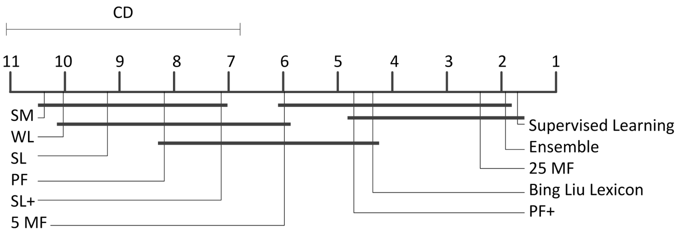

Fig. 2. Comparison of all sentiment polarity methods against each other with the Nemenyi post-hoc test using f-measure results. Groups of methods that are not significantly different (at $p = 0.05$) are connected.

TABLE III. RESULTS FOR EACH CORPUS OBTAINED BY LEXICON-BASED ENSEMBLE AND SUPERVISED LEARNING APPROACHES.

Corpus-Approach	Accuracy	Precision	Recall	F-measure
Automotive-Lex.Ensem.	0.485	0.473	0.485	0.469
Automotive-Sup.Learn.	0.501	0.502	0.501	0.501
Book-Lex.Ensem.	0.505	0.501	0.505	0.499
Book-Sup.Learn.	0.478	0.478	0.478	0.478
Electronics-Lex.Ensem.	0.501	0.49	0.501	0.491
Electronics-Sup.Learn.	0.468	0.469	0.468	0.468
Health-Lex.Ensem.	0.487	0.489	0.487	0.486
Health-Sup.Learn.	0.509	0.51	0.509	0.509
Movies-Lex.Ensem.	0.504	0.493	0.504	0.491
Movies-Sup.Learn.	0.496	0.497	0.496	0.496

2) Comparison of Lexicon-based Sentiment Result: The analysis of classification results obtained by Lexicon-based approach that used individually each of nine lexicons (with no ensemble fusion) revealed that there are two separable groups of lexicons that have produced statistically distinct results. As it can be seen in Table V-VIII the 25MF, Bing Liu Opinion Lexicon (BL) and PF+ lexicons had better predictive capability than the rest. Moreover, the 25 MF lexicon, obtained by newly proposed method, resulted with better overall rank and classification accuracy measures across all corpuses in comparison to Bing Liu's lexicon.

3) Time Complexity of Lexicon-based Ensemble vs. Supervised Learning: The supervised learning approach was much slower than the ensemble method (Table IV). It can be seen that Lexicon-based Ensemble is faster by more than two orders of magnitude in comparison to Supervised Learning method. Due to the fact that these methods share similar accuracy, the computational complexity shows that BoW ensemble method is reasonable choice and proved to be efficient and fast across different corpuses.

V. CONCLUSIONS AND FUTURE WORK

The paper introduced a new approach to lexicon extraction that can be successfully used for sentiment polarity assignment. It has been shown that accuracy obtained from such lexicons outperforms other lexicon based approaches. Moreover,

TABLE II. LEXICONS USED IN RESEARCH.

Lexicon	Positive words	Negative words
SM	good	bad
SL	good, awesome, great, fantastic, wonderful	bad, terrible, worst, sucks, awful, dumb
SL+	good, awesome, great, fantastic, wonderful, best, love, excellent	bad, terrible, worst, sucks, awful, dumb, waist, boring, worse
PF	will, has, must, is	was, would, had, were
PF+	will, has, must, is, good, awesome, great, fantastic, wonderful, best, love, excellent	was, would, had, were, bad, terrible, worst, sucks, awful, dumb, waist, boring, worse
WL	perfection, captures, wonderfully, powell, refreshing, flynn, delightful, gripping, beautifully, underrated, superb, delight, welles, unforgettable, touching, favorites, extraordinary, stewart, brilliantly, friendship, wonderful, magnificent, marie, jackie	horrible, unconvincing, uninteresting, insult, uninspired, sucks, miserably, boredom, cannibal, godzilla, lame, wasting, remotely, awful, poorly, laughable, worst, lousy, redeeming, atrocious, pointless, pointless, blah, waste, unfunny, seagal
Bing Liu Lexicon	2006 words	4783 words
5 MF Automotive	perfect, perfectly, exactly, easy, happy	company, wrong, return, returned, sent
25 MF Automotive	heavy, best, perfect, nice, bike, easy, recommend, perfectly, happy, run, clean, power, gas, price, exactly, overall, job, fits, easily, great, works, installation, filter, install, stock	ordered, money, year, returned, try, said, pay, tried, weeks, correct, sent, picture, return, description, company, completely, trying, disappointed, wont, received, maybe, months, wrong, thought, model
5 MF Books	loved, excellent, economics, letter, highly	poor, money, waste, disappointed, boring
25 MF Books	beautiful, love, simple, excellent, highly, human, best, perfect, living, ways, wonderful, easy, today, gives, government, understanding, letter, child, must, loved, great, clear, economics, introduction, hope	money, lack, edition, buy, trying, version, waste, poor, finish, else, completely, words, nothing, believe, disappointed, pages, word, unfortunately, anything, maybe, reviews, bad, either, boring, example
5 MF Electronics	comfortable, love, pleased, bed, excellent	warranty, customer, stopped, return, poor
25 MF Electronics	love, installed, pleased, best, size, perfect, built, fast, quickly, easy, recommend, perfectly, happy, far, price, excellent, comfortable, great, room, charger, bed, air, bag, works, original	warranty, paper, tried, service, pay, support, send, stopped, sent, told, poor, return, returned, started, company, phone, broke, nothing, disappointed, customer, unfortunately, months, tech, went, piece
5 MF Health	dry, feet, love, best, highly	money, flavor, waste, disappointed, company
25 MF Health	love, years, feet, skin, best, things, amazing, long, recommended, definitely, easy, recommend, happy, pain, goes, every, wear, foot, highly, dry, great, stuff, night, works, husband	ordered, battery, money, cheap, second, flavor, rather, taste, instead, waste, smell, gave, read, company, felt, hold, disappointed, buying, received, unfortunately, batteries, thought, bad, either, blades
5 MF Movies	highly, season, amazing, wonderful, excellent	waste, worst, terrible, worse, horrible
25 MF Movies	beautiful, heart, liked, love, concert, family, son, home, enjoyed, best, perfect, amazing, wonderful, war, season, price, excellent, highly, true, loved, great, always, definitely, husband, songs	absolutely, money, rent, annoying, tried, script, except, horrible, save, poor, worse, horror, wrong, worst, nothing, scary, waste, unfortunately, boring, terrible, bad, stupid, supposed, completely, minutes

TABLE IV. COMPARISON OF EXECUTION TIME [IN SECONDS] FOR EACH CORPUS OF LEXICON-BASED ENSEMBLE AND SUPERVISED LEARNING APPROACHES.

Corpus	Lex.Ensem. [s]	Sup.Learn. [s]	Ratio [Sup.Learn./Lex.Ensem.]
Automotive	54	12 329	228
Books	89	21 324	240
Electronics	69	13 863	201
Health	52	13 191	254
Movies	97	23 673	244

TABLE V. ACCURACY OF PARTICULAR LEXICON APPROACHES FOR EACH CORPUS.

Dataset	SM	SL	SL+	PF	PF+	WL	BL*	5 MF	25 MF
Automotive	0,34	0,36	0,37	0,34	0,38	0,36	0,42	0,42	0,47
Books	0,30	0,35	0,38	0,35	0,40	0,39	0,43	0,42	0,48
Electronics	0,37	0,36	0,38	0,34	0,38	0,36	0,43	0,42	0,50
Health	0,33	0,37	0,39	0,37	0,42	0,36	0,42	0,42	0,47
Movies	0,33	0,39	0,40	0,36	0,41	0,42	0,45	0,42	0,49

* BL - Bing Liu Lexicon

TABLE VI. PRECISION OF PARTICULAR LEXICON APPROACHES FOR EACH CORPUS.

Dataset	SM	SL	SL+	PF	PF+	WL	BL	5 MF	25 MF
Automotive	0,38	0,43	0,44	0,34	0,39	0,50	0,43	0,53	0,48
Books	0,34	0,42	0,45	0,36	0,40	0,54	0,42	0,53	0,48
Electronics	0,37	0,43	0,45	0,34	0,38	0,48	0,46	0,54	0,51
Health	0,37	0,42	0,44	0,37	0,43	0,53	0,43	0,53	0,47
Movies	0,39	0,45	0,47	0,36	0,41	0,54	0,44	0,55	0,50

TABLE VII. RECALL OF PARTICULAR LEXICON APPROACHES FOR EACH CORPUS.

Dataset	SM	SL	SL+	PF	PF+	WL	BL	5 MF	25 MF
Automotive	0,34	0,36	0,37	0,34	0,38	0,36	0,42	0,42	0,47
Books	0,30	0,35	0,38	0,35	0,40	0,39	0,43	0,42	0,48
Electronics	0,33	0,36	0,38	0,34	0,38	0,36	0,43	0,42	0,50
Health	0,33	0,37	0,39	0,37	0,42	0,36	0,42	0,42	0,47
Movies	0,33	0,39	0,40	0,36	0,41	0,42	0,45	0,42	0,49

TABLE VIII. F-MEASURE OF PARTICULAR LEXICON APPROACHES FOR EACH CORPUS.

Dataset	SM	SL	SL+	PF	PF+	WL	BL	5 MF	25 MF
Automotive	0,25	0,30	0,32	0,32	0,38	0,22	0,40	0,38	0,47
Books	0,24	0,31	0,35	0,33	0,40	0,32	0,40	0,38	0,48
Electronics	0,26	0,32	0,34	0,32	0,38	0,23	0,40	0,38	0,50
Health	0,25	0,32	0,34	0,35	0,42	0,24	0,40	0,39	0,46
Movies	0,29	0,37	0,38	0,34	0,41	0,37	0,42	0,37	0,49

we have shown that simplistic BoW methods with ensemble classifiers are much faster than a supervised approach to sentiment classification while yielding similar accuracy. BoW methods also are proved efficient and fast across all datasets. We thus conclude that such methods should be more widely used when available computational time is limited. Nevertheless, the presented methods still do not satisfy expectations and more complex sentiment analysis aspects such as proposed by Socher et al.,[16] should be explored in order to achieve higher accuracy.

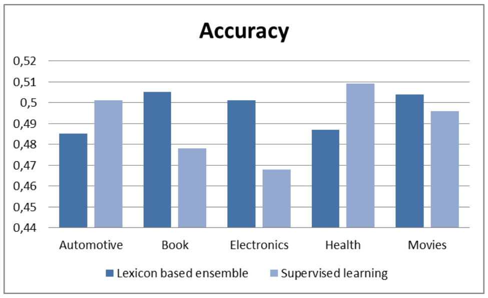

Fig. 3. Classification accuracy of Lexicon-based Ensemble and Supervised Learning approaches.

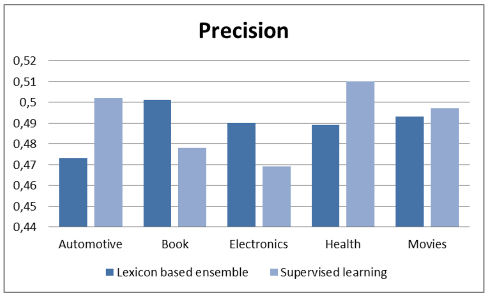

Fig. 4. Comparison of precision of Lexicon-based Ensemble and Supervised Learning approaches.

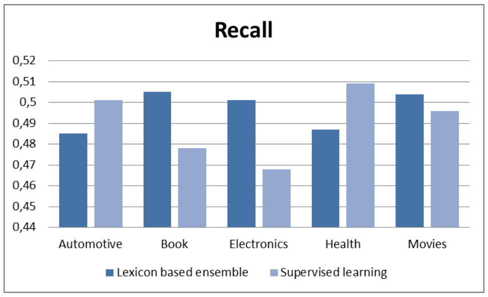

Fig. 5. Recall comparison of Lexicon-based Ensemble and Supervised Learning approaches.

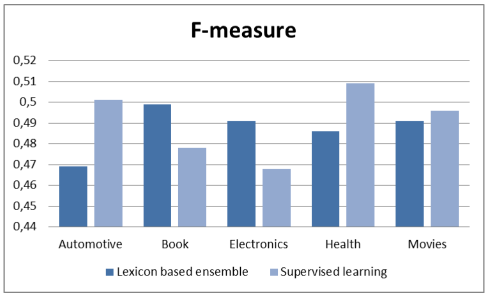

Fig. 6. Comparison of F-measure of Lexicon-based Ensemble and Supervised Learning approaches.

ACKNOWLEDGMENT

The work was partially supported by The National Science Centre, decision no. DEC-2013/09/B/ST6/02317 and the European Commission under the 7th Framework Programme, Coordination and Support Action, Grant Agreement Number 316097, Engine project.

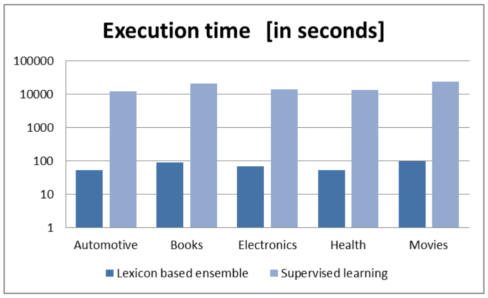

Fig. 7. Execution time of Lexicon-based Ensemble and Supervised Learning approaches.

REFERENCES

[1] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis." in *Mining Text Data*. Springer, 2012, pp. 415–463.

[2] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," pp. 174–181, 1997.

[3] S.-H. Na, Y. Lee, S.-H. Nam, and J.-H. Lee, "Improving opinion retrieval based on query-specific sentiment lexicon," in *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science, vol. 5478. Springer Berlin / Heidelberg, 2009, pp. 734–738.

[4] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1–135, Jan. 2008.

[5] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," pp. 417–424, 2002.

[6] D. Maynard and M. A. Greenwood, "Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis," 2014.

[7] D. Preotiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins, and M. Niranjan, "Trendminer: An architecture for real time analysis of social media text," in *Sixth International AAAI Conference on Weblogs and Social Media*, 2012, pp. 38–42.

[8] M. Whitehead and L. Yaeger, "Sentiment mining using ensemble classification models." in *SCSS (1)*. Springer, 2008, pp. 509–514.

[9] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[10] R. E. Schapire, "A brief introduction to boosting," in *IJCAI*, 1999, pp. 1401–1406.

[11] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[12] L. Augustyniak, T. Kajdanowicz, P. Kazienko, M. Kulisiewicz, and W. Tuliglowicz, "An approach to sentiment analysis of movie reviews: Lexicon based vs. classification," in *HAIS*, 2014, pp. 168–178.

[13] B. Jason, "Practical sentiment analysis tutorial," 2014, sentiment Analysis Symposium.

[14] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 142–150.

[15] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text," in *The 7th ACM conference on Recommender systems*. ACM, 2013, pp. 165–172.

[16] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, 2013, pp. 1631–1642.