ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 6, Nomor 3, Desember 2017

METODE LEXICON-LEARNING BASED UNTUK IDENTIFIKASI TWEET OPINI BERBAHASA INDONESIA

Yufis Azhar

Teknik Informatika, Universitas Muhammadiyah Malang
Malang, Indonesia

e-mail: yufis@umm.ac.id

Abstrak

Media sosial telah lama digunakan masyarakat untuk menyampaikan opini maupun fakta terhadap suatu kejadian, khusunya twitter. Banyak metode yang diusulkan untuk mengekstrak tweet yang berisi opini. Diantaranya mengunakan pendekatan identifikasi kata kunci dalam suatu tweet yang lebih dikenal dengan istilah lexicon based. Meskipun metode ini memiliki nilai presisi yang cukup tinggi dalam mengidentifikasi suatu tweet opini, akan tetapi nilai recall yang dihasilkan cukup rendah. Hal ini karena keterbatasan lexicon yang digunakan sebagai identifier. Dalam penelitian ini, diusulkan kombinasi metode lexicon based dan machine learning dalam mengoptimalkan hasil identifikasi tweet opini. Hasil pengujian menunjukkan peningkatan nilai recall yang cukup signifikan jika dibandingkan dengan metode lexicon based dengan tetap menjaga nilai precision.

Kata kunci: tweet opini, lexicon based, learning based

Abstract

Social media has long used the community to convey opinions and facts about an event, especially twitter. Many methods are proposed to extract tweets containing opinions. Among them using a keyword identification approach in a tweet better known as lexicon based. Although this method has a high enough precision value in identifying an opinion tweet, the resulting recall value is quite low. This is because of the limitations of lexicon used as identifier. In this research, proposed combination of lexicon based and machine learning method in optimizing the result of identification of opinion tweet. The test results showed a significant increase in recall value when compared with lexicon based method while maintaining precision value

Keywords : *opinion tweet, lexicon based, learning based*

PENDAHULUAN

Sebagian orang lebih suka menyampaikan pendapat (opini) melalui dunia maya (internet). Baik itu lewat blog, forum online, media sosial (seperti facebook, twiiter atau instagram), maupun melalui situs-situs yang memang menyediakan fitur untuk user bisa berkomentar (seperti situs berita, situs jual beli, dan sebagainya). Banyaknya data opini berupa teks yang tersebar di internet menjadi daya tarik bagi sebagian peneliti untuk memanfaatkan data tersebut. Sebab opini-opini ini bisa berguna untuk membaca tingkat kepuasan konsumen atau masyarakat terhadap suatu produk maupun kebijakan pemerintah. Bidang ilmu yang fokus mempelajari masalah ini disebut opinion mining.

Meskipun penelitian tentang opinion mining untuk tweet berbahasa Indonesia cukup banyak dilakukan [1][2], akan tetapi sebagian besar masih diperuntukkan untuk klasifikasi polaritas dari suatu opini (apakah opini positif atau negatif). Padahal untuk dapat mengklasifikasikan suatu opini, seorang peneliti harus yakin terlebih dahulu bahwa teks tersebut benar-benar mengandung opini. Dari penelitian yang telah dilakukan sebelumnya [3], didapatkan kesimpulan bahwa proses identifikasi apakah suatu teks mengandung opini atau

tidak memiliki peran yang sangat penting dalam pengukuran akurasi suatu metode.

Ada 2 cara yang umumnya digunakan oleh peneliti dalam menentukan apakah suatu teks mengandung kalimat opini atau tidak. Cara pertama adalah dengan membaca isi teks secara keseluruhan, kemudian memberikan label secara manual. Cara ini memiliki presisi yang cukup tinggi tapi sangat sulit diterapkan untuk data yang berjumlah sangat besar karena memakan banyak waktu. Sedangkan cara yang kedua adalah dengan menerapkan metode yang dapat mengekstrak kalimat opini secara otomatis. Metode ekstraksi yang umum dipakai adalah *lexicon-based* [4][5] dan *learning-based* [6][7].

Metode *lexicon-based* bekerja dengan cara membuat kamus kata opini (*lexicon)* terlebih dahulu. Kata-kata yang terdapat pada kamus tersebut digunakan untuk mengidentifikasi apakah suatu kalimat mengandung opini atau tidak. Sedangkan metode *learning-based* memanfaatkan *machine learning*. Metode ini akan mengklasifikasikan teks opini secara otomatis dengan melihat data latih berupa teks opini yang sudah diklasifikasikan secara manual sebelumnya.

Metode *lexicon-based* memang dapat mengekstrak kalimat opini dengan presisi yang sangat tinggi. Akan tetapi, nilai *recall* nya rendah. Sedangkan metode *learning-based* memiliki nilai presisi dan recall cukup tinggi, tetapi sangat bergantung pada jumlah data latih. Metode ini harus menggunakan data latih yang cukup banyak agar dapat menghasilkan model *classifier* yang baik. Dalam penelitian ini, diusulkan kombinasi dari kedua metode tersebut. Diharapkan dengan mengkombinasikan keduanya hasil klasifikasi yang didapat memiliki nilai presisi dan *recall* yang tinggi meskipun data latih yang digunakan sedikit.

METODE YANG DIUSULKAN

Pada masa lalu, sebagian besar peneliti di bidang opinion mining hanya terfokus pada dataset berupa dokumen review yang susunan bahasanya baku. Seiring dengan perkembangan media sosial seperti facebook dan twitter, opini tidak lagi hanya terdapat pada forum online, blog ataupun situs jual beli. Media sosial seperti twitter menjelma menjadi buku harian publik yang banyak berisi opini dari penggunanya tentang berbagai hal. Oleh karena itu, banyak peneliti yang kemudian beralih melakukan analisa opini di media sosial khusunya twitter (Hu, 2013; Kouloumpis, 2011; Pak, 2010, Nakov, 2013). Dalam penelitian ini, twitter juga digunakan sebagai tempat untuk mendapatkan dataset. Twitter dipilih karena ketersedian API yang cukup lengkap fitur-fiturnya untuk keperluan crawling data.

Mengolah data tweet memiliki tantangan tersendiri. Berbeda dengan dokumen review yang memiliki struktur bahasa baku, banyak tweet yang memiliki struktur dan gaya bahasa tidak baku. Hal ini menyulitkan dalam tahap preprocessing seperti tokenizing, stemming, parsing ataupun POS Tagging. Terlebih untuk tweet berbahasa Indonesia yang memiliki puluhan bahasa daerah yang kadangkala tercampur dalam suatu tweet. Selain itu, tools yang tersedia untuk pengolahan NLP dalam bahasa Indonesia tidak sebanyak jika dibandingkan dengan bahasa Inggris. Hal ini membuat penelitian yang menggunakan tweet berbahasa Indonesia tidak sebanak penelitian sejenis dengan menggunakan bahasa Inggris. Beberapa penelitian terkait opinion mining pada tweet berbahasa Indonesia yang pernah dilakukan antara lain (Calvin, 2014; Aliandu, 2014; Wicaksono, 2014).

Liu (2007) mendefinisikan bahwa dalam opinion mining, terdapat 2 sub-pekerjaan utama, yakni klasifikasi subjektifitas dan sentiment analysis (klasifikasi opini). Maksud dari klasifikasi subjektifitas adalah membedakan antara kalimat opini dan non-opini, sedangkan klasifikasi opini bertujuan membedakan antara opini positif dan negatif. Dari hasil penelitian yang telah dilakukan sebelumnya, dapat disimpulkan bahwa klasifikasi subjektifitas memiliki peran yang sangat penting untuk menjaga akurasi sentiment analysis tetap tinggi (Azhar, 2016). Ada 2 cara yang umumnya

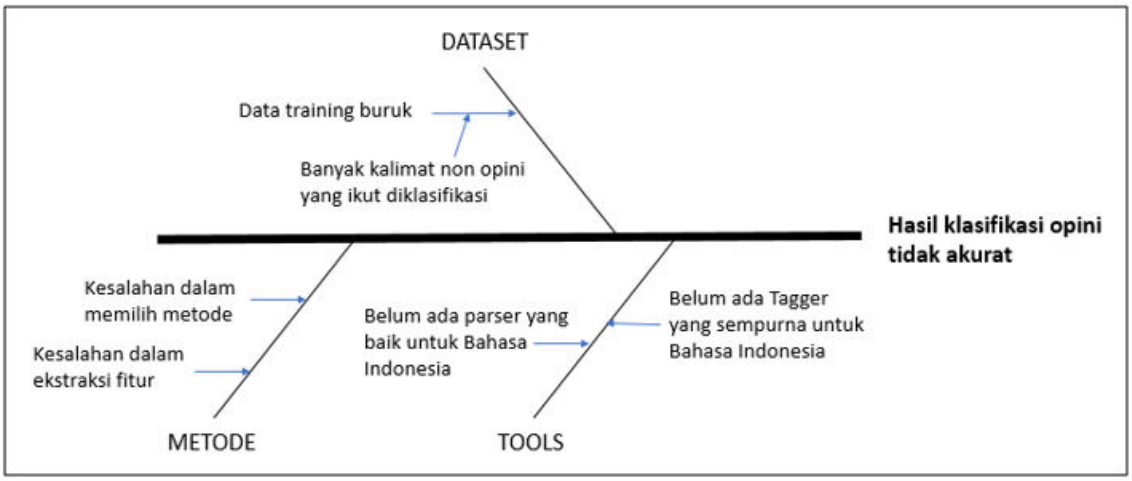
digunakan oleh peneliti dalam menentukan apakah suatu teks mengandung kalimat opini atau tidak. Cara pertama adalah dengan membaca isi teks secara keseluruhan, kemudian memberikan label secara manual. Cara ini memiliki presisi yang cukup tinggi tapi sangat sulit diterapkan untuk data yang berjumlah sangat besar karena memakan banyak waktu. Sedangkan cara yang kedua adalah dengan menerapkan metode yang dapat mengekstrak kalimat opini secara otomatis. Metode ekstraksi yang umum dipakai adalah lexicon-based (Pak, 2010; Ding, 2008; Taboada, 2010; Azhar, 2013), rule-based (Rozi, 2013; Azhar, 2016), dan learning-based (Pang, 2002; Azhar, 2015). Metode lexicon-based bekerja dengan cara membuat kamus kata terlebih dahulu. Kata-kata yang terdapat pada kamus tersebut digunakan untuk mengidentifikasi apakah suatu kalimat mengandung opini atau tidak. Sedangkan metode learning-based memanfaatkan machine learning. Metode ini akan mengklasifikasikan teks opini secara otomatis dengan melihat data latih yang berupa teks opini yang sudah diklasifikasikan secara manual sebelumnya.

Pada tahun 2013, telah dilakukan penelitian tentang otomatisasi perbandingan produk berdasarkan bobot fitur pada teks opini (Azhar, 2013). Pada penelitian ini, diusulkan metode Modification Double Propagation (MDP) untuk melakukan perbandingan pada dua produk berdasarkan review dari customer secara otomatis. Dalam penelitian ini, hasil yang didapat adalah metode yang diusulkan dapat dengan baik memberikan scoring terhadap suatu produk berdasarkan polaritas opininya. Akan tetapi nilai recall untuk ekstraksi target opininya kurang baik.

Pada tahun 2014, dilakukan penelitian lanjutan untuk menyempurnakan hasil ekstraksi target opini pada penelitian sebelumnya (Azhar, 2014). Dalam penelitian ini, kata ganti (pronoun) juga diperhatikan sehingga diharapkan meningkatkan nilai recall dari penelitian sebelumnya.

Sementara penelitian yang dilakukan di tahun 2015, bertujuan untuk meningkatkan nilai precision hasil ekstraksi terget opini dari kedua penelitian sebelumnya (Azhar, 2015). Caranya adalah dengan mengklasifikasikan target opini berdasarkan kedekatannya terhadap keyword yang sudah ditentukan sebelumnya.

Ketiga penelitian tersebut menggunakan dataset berbahasa Inggris yang diambil dari dokumen review suatu produk yang dijual melalui situs amazon.com. Pada tahun 2016, peneliti mencoba menggunakan dataset berbahasa Indonesia yang diambil dari media social twitter (Azhar, 2016). Dalam penelitian ini, metode-metode yang sudah dikembangkan pada penelitian sebelumnya digunakan. Hasilnya system dapat mengklasifikasikan tweet dengan baik. Akan tetapi, jika melihat nilai precision dan recall dari tiap kelas yang dihasilkan, ternyata nilainya tidak cukup tinggi. Hal ini menjadi catatan bagi peneliti



Gambar 1. Fishbone Diagram Untuk Menemukan Akar Permasalahan

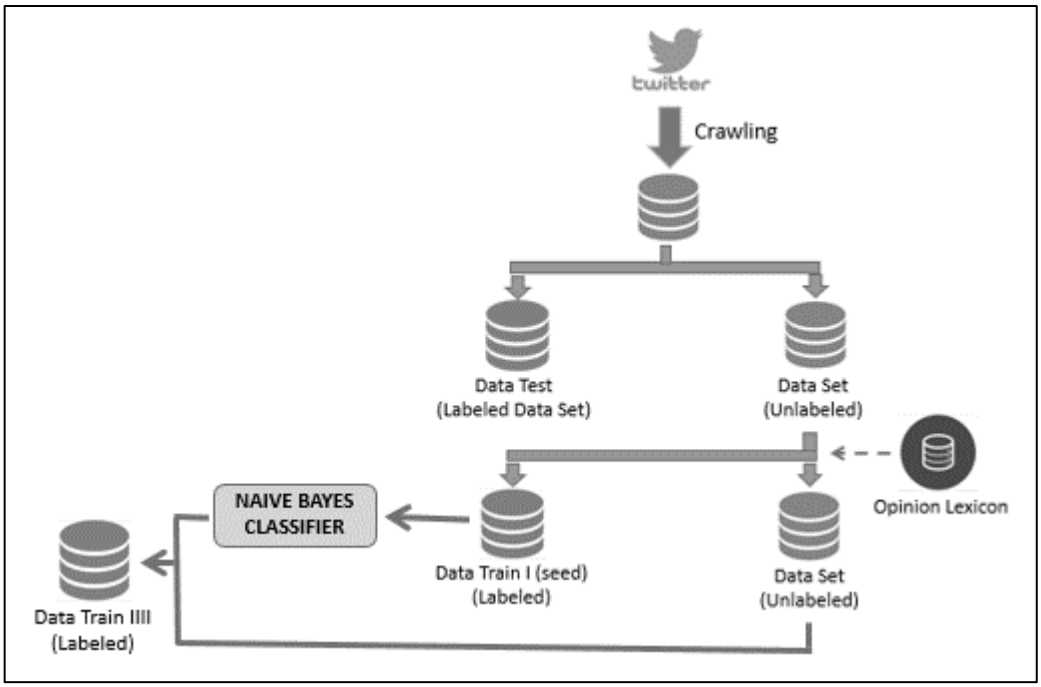
untuk disempurnakan pada penelitian berikutnya.

Berdasarkan hasil penelitian yang telah dilakukan sebelumnya, peneliti berusaha menemukan akar permasalahan mengapa hasil klasifikasi opini tidak akurat, padahal di beberapa penelitian sebelumnya, dengan menggunakan metode yang sama, yang menggunakan dataset berbahasa Inggris, hasil klasifikasinya cukup bagus. Dari hasil Analisa, dibuatlah fishbone diagram seperti Gambar 1. Salah satu factor yang dapat mengakibatkan hasil klasifikasi buruk adalah kondisi dataset yang juga buruk. Banyak tweet yang sebenarnya bukan merupakan kalimat opini, ikut dijadikan sebagai data latih untuk mengklasifikasikan opini positif dan negatif. Hal tersebut tentunya dapat mempengaruhi akurasi hasil klasifikasi.

Oleh karena itu, dalam penelitian ini diusulkan suatu metode untuk membuat data l atih yang hanya memuat tweet opini saja secara otomatis. Kombinasi dari lexicon based, rule based, dan learning based digunakan untuk meningkatkan nilai precision dan recall dari hasil klasifikasi yang dihasilkan.

Secara umum, terdapat 4 tahapan dalam penelitian ini, yaitu (1) Pengumpulan dataset dari twitter; (2) Pembuatan data test; (3) Pembuatan data train; serta (4) Evaluasi classifier. Untuk lebih jelasnya dapat dilihat pada gambar alur kerja system pada Gambar 2.

Pada penelitian ini, dataset didapatkan dengan cara *crawling tweet* berbahasa Indonesia berdasarkan *keyword* menggunakan *Twitter API*. *Keyword* yang dipilih berupa produk, perusahaan ataupun tokoh publik yang sering dijadikan sasaran opini. Dari dataset yang telah didapatkan, kemudian diambil 5% nya sebagai data test. Data test ini kemudian akan diklasifikasikan oleh 2 orang anotator secara manual. Kedua anotator ini bekerja secara *independent* untuk melabeli data test sebagai tweet opini dan tweet non-opini. Data hasil anotator kemudian digabungkan. Jika kedua anotator setuju tentang jenis label suatu tweet, maka tweet tadi dimasukkan ke dalam list data test. Akan tetapi, jika jenis label dari suatu tweet berbeda, maka anotator ke-3 diminta untuk melabeli tweet tersebut.



Gambar 2. Alur Kerja Sistem

Data yang tersisa setelah pengambilan data test, selanjutnya akan digunakan sebagai data train. Dalam penelitian ini, diusulkan 2 tahapan pembuatan data train. Yakni *lexicon-based*, dan *learning-based*.

a) Lexicon-based

Tahap pertama dalam pembuatan data train dimulai dengan pendekatan lexicon-based. Pada metode ini, setiap tweet yang ada pada dataset akan dianalisa satu per satu. Tweet yang mengandung kata yang terdapat pada kamus kata opini akan dilabeli sebagai tweet opini. Kamus kata opini yang dipergunakan dalam penelitian ini didapatkan dari penelitian yang telah dilakukan sebelumnya [3]. Kamus kata tersebut merupakan versi terjemah dari senti-wordnet yang telah dimodifikasi disesuaikan dengan karakter user di Indonesia, ditambah dengan karakter-karakter *emoticon* yang sering digunakan oleh user ketika menyampaikan opini. Hasil pelabelan dari metode ini kemudian dinamakan Data Train I.

(b) Learning-based

Metode lexicon-based memang mampu mengekstrak kalimat opini dengan nilai presisi yang cukup tinggi, akan tetapi nilai recall nya terbilang rendah [3]. Sehingga tidak semua kalimat opini mampu diekstrak oleh kedua metode tersebut. Oleh karena itu, di penelitian ini juga digunakan pendekatan *learning-based* untuk melabeli sisa dataset yang belum terlabeli oleh metode sebelumnya. Metode klasifikasi yang dipilih adalah naive bayes karena terbukti cukup baik digunakan untuk keperluan sentiment analysis [2].

Karena metode ini adalah jenis *supervised learning*, maka untuk dapat membuat model classifiernya dibutuhkan data train sementara. Data Train I yang berisi tweet opini (hasil dari metode sebelumnya) digunakan sebagai data train dengan class opini. Sedangkan untuk data train kelas non-opininya, dilakukan analisa sederhana terhadap dataset yang belum terlabeli. Tweet yang berasal dari akun situs berita, atau lembaga pemerintah dilabeli sebagai tweet non-opini. Hal ini dikarenakan tweet milik akun-akun tersebut biasanya berisi informasi deskriptif, bukan opini. Kedua class tersebut kemudian digunakan sebagai data train untuk melabeli tweet-tweet lain dengan menggunakan algoritma naive-bayes. Hasil labelisasi inilah yang nantinya dinamakan sebagai Data Train II.

Hasil akhir dari tahap sebelumnya adalah Data Train II yang berisi semua dataset (non data test) yang telah terlabeli sebagai tweet opini atau tweet non-opini. Selanjutnya data train ini akan digunakan untuk menguji performa sistem pada data test. Metode Naive Bayes kembali digunakan untuk melakukan klasifikasi. Nilai yang akan dilihat dari proses klasifikasi ini adalah *success_rate, precision*, serta *recall*. Nilai ketiganya kemudian akan dibandingkan dengan hasil klasifikasi yang data trainnya disusun dengan menggunakan metode lain.

HASIL DAN PEMBAHASAN

Dataset dikumpulkan dengan metode crawling menggunakan twitter API. Dari proses *crawling* yang dilakukan selama kurang lebih 1 minggu, didapatkan 1.532.879 tweet. Kemudian, tweet yang sudah dikupulkan tersebut difilter dengan cara membuang tweet yang memenuhi kriteria: (a) Tweet merupakan retweet dari tweet lain yang ada di dataset; (b) Tweet hanya berisi tautan ke situs lain; (c) Tweet hanya berisi gambar.

Dari proses filtering tersebut, didapatkan 1.004.535 tweet yang tersisa. Data tersebut kemudian dibagi menjadi 2 bagian. Bagian pertama berisi 50.000 tweet yang nantinya akan digunakan sebagai data uji, dan bagian kedua berisi 954.535 tweet yang akan digunakan sebagai data training.

Data uji kemudian akan dilabeli secara manual oleh 2 orang annotator. Jika kedua annotator berbeda pendapat tentang labelisasi suatu tweet, maka tweet tersebut diberian pada annotator ke-3 untuk dilabeli. Dari proses ini, didapatkan 17.436 tweet yang dilabeli sebagai tweet opini dan 32.564 tweet dilabeli sebagai non-opini.

Data sisa sebanyak 954.535 yang belum terlabeli kemudian akan dilabeli secara otomatis dengan menggunakan metode *lexicon-based*. Tweet yang mengandung kata kunci akan dilabeli sebagai tweet opini. Dari proses ini, didapatkan 153.786 tweet yang terekstak sebagai tweet opini. Sedangkan sisanya sebesar 800.749 tweet dianggap sebagai tweet non-opini.

Tweet yang terdapat dalam class opini kembali diklasifikasikan. Kali ini dengan menggunakan metode machine learning, yaitu Naïve Bayes. Untuk dapat melakukannya, data yang terdapat pada class opini (hasil dari langkah sebelumnya) sebanyak 153.786 tweet digunakan sebagai data latih untuk class opini. Sedangkan untuk class non-opini, dilakukan langkah sederhana untuk mengekstrak tweet yang dianggap sebagai fakta. Tweet yang berasal dari akun situs berita, atau lembaga pemerintah dilabeli sebagai tweet non-opini. Hal ini dikarenakan tweet milik akun-akun tersebut biasanya berisi informasi deskriptif, bukan opini. Kedua class tersebut kemudian digunakan sebagai data train untuk melabeli tweet-tweet lain dengan menggunakan algoritma naive-bayes. Dari proses ini, didapatkan 245.562 tweet opini. Sehingga jika digabungkan dengan hasil dari 2 proses sebelumnya, didapatkan hasil seperti yang dapat dilihat pada Tabel 1.

Tabel 1. Hasil Ekstraksi Tweet

Metode	Jumlah Dataset	Class Opini	Class Non Opini
Lexicon Based	954,535	153,786	800,749
Learning-Based	800,749	245,562	555,187
TOTAL		**399,348**	**555,187**

Tabel 2. Hasil Pengujian

Ukuran Uji	Lex	Lex-Learn
Success_Rate	0.92	0.94
Recall	0.78	0.88
Precision	0.86	0.87

Data tersebut kemudian digunakan sebagai data latih untuk menguji data testing. Dari proses pengujian yang dilakukan, didapatkan hasil seperti pada Tabel 2.

Dari kedua Tabel tersebut, dapat dilihat bahwa metode yang diusulkan terbukti dapat meningkatkan nilai recall dibandingkankan metode sebelumnya dengan peningkatan sebesar 11%. Peningkatan recall ini diimbangi juga dengan tetap menjaga nilai precision yang berada di kisaran angka 87%.

KESIMPULAN

Dari pengujian dapat dilihat bahwa algoritma yang diusulkan telah mampu meningkatkan nilai recall dari metode lexicon based dengan tetap mempertahankan nilai precision dari metode tersebut.

REFERENSI

[1] Setiawan J. Using Text Mining to Analyze Mobile Phone Provider Service Quality (Case Study: Social Media Twitter). International Journal of Machine Learning and Computing. 2014 Feb 1;4(1):106.
[2] Aliandu P. Sentiment analysis on indonesian tweet. The Proceedings of The 7th ICTS. 2014.
[3] Azhar, Y. Klasifikasi Fitur Dalam Dokumen Review Produk Dengan Metode Local Pointwise Mutual Information. Network Engineering Research Operation [NERO]. 2016. 2(1).
[4] Kiritchenko S, Zhu X, Mohammad SM. Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research. 2014;50:723-62.
[5] Kontopoulos E, Berberidis C, Dergiades T, Bassiliades N. Ontology-based sentiment analysis of twitter posts. Expert systems with applications. 2013 Aug 31;40(10):4065-74.
[6] Smailović J, Grčar M, Lavrač N, Žnidaršič M. Stream-based active learning for sentiment analysis in the financial domain. Information sciences. 2014 Nov 20;285:181-203.
[7] Neethu MS, Rajasree R. Sentiment analysis in twitter using machine learning techniques. InComputing,

Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on 2013 Jul 4 (pp. 1-5). IEEE