이재용 선처 여론이 60%? '빅데이터 구라'의 실체

지윤성 팩트체커 | 승인 2020.06.09 10:40

> 엉터리 빅데이터 분석, 이를 여과없이 인용하는 한국 언론

2020년 6월 8일 이재용 삼성전자 부회장의 구속영장 실질심사를 앞두고 **조선일보**를 포함한 언론들은 일제히 '빅데이터가 본 국민생각'이라는 타이틀로 "이재용 삼성전자 부회장에 대한 검찰의 구속영장 청구에 대해 국민들은 '불관용'보다는 **내심** '선처'를 더 바라고 있는 것으로 나타났다"고 밝혔다. <글로벌빅데이터연구소>의 빅데이터 분석결과라는 꼬리표를 달았다. **내심**이라는 비과학적인 용어에서부터 의심이 들었다. 왜 슬픈 예감은 틀리지 않는 것일까.

<글로벌빅데이터연구소>라는 곳의 **원문**을 살펴보니 관련 업계 종사자로서 비참함마저 들었다. 연구방법론에 대한 구체적인 설명도 없고 "단어 기반 연관어 분석기법이 빅데이터상 국민들의 의견을 분석할 때 가장 유용한 방법 중 하나"라는 근거없는 표현이 적혀있다. 이 분야에서 중요시하는 데이터 샘플링 편향에 대한 주의도 없다. <글로벌빅데이터연구소>라는 곳의 발표자료의 문제점을 하나하나 살펴보자.

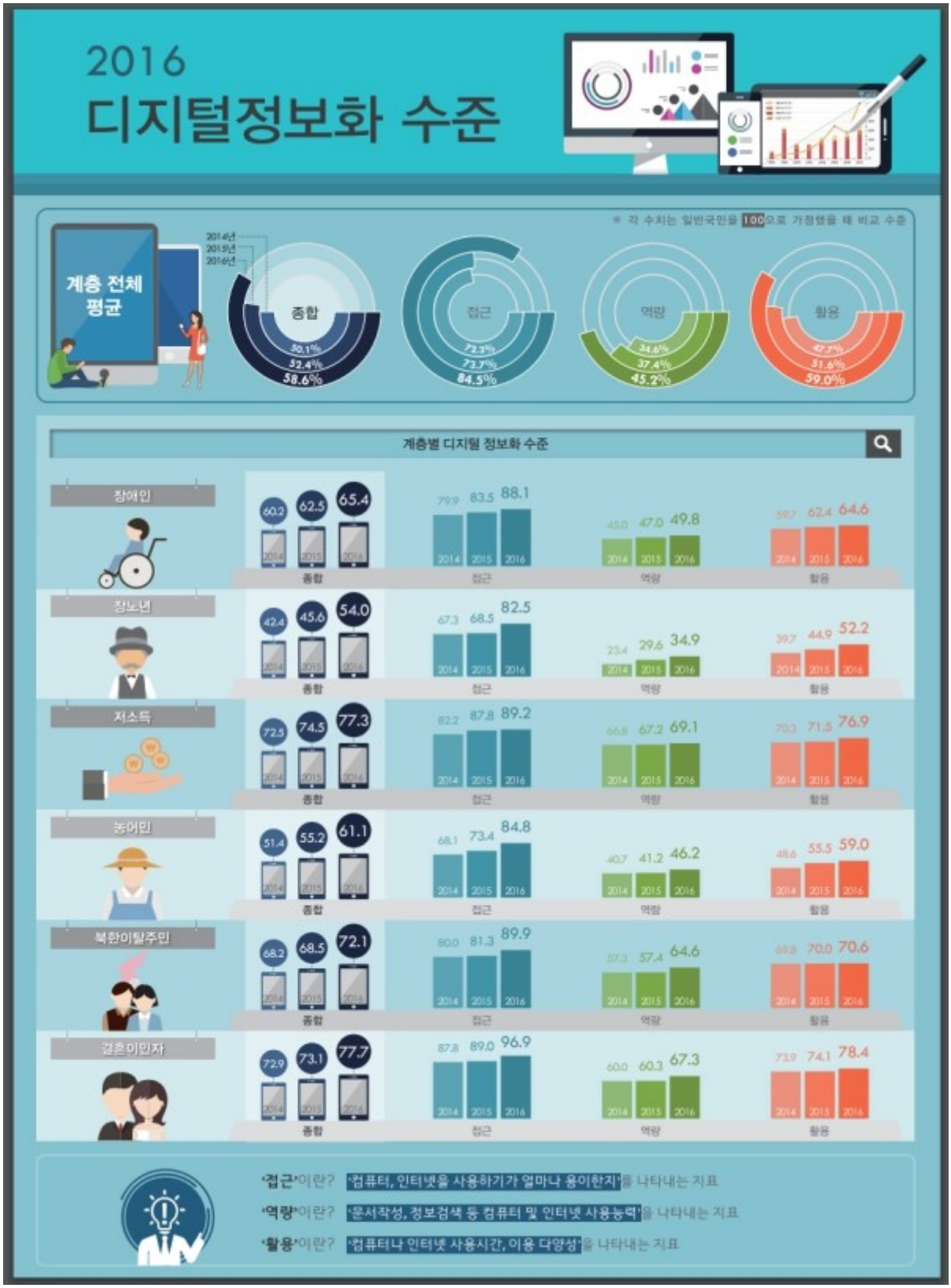
"이재용 선처가 60%" 기사가 넘쳐나지만 누구도 데이터의 신뢰도를 조사하지 않았다.

1. 아무거나 긁어오면 여론? 데이터 출처가 명확해야 한다

빅데이터 분석엔 항상 편향성을 감안해야 한다. 정보 격차로 인해 특정 계층에 데이터 생산이 편중되어 있기 때문에 자료 수집시 그 계층이 과대대표될 가능성이 있다. 데이터 마이닝(샘플링) 단계에서부터 주의가 요구되는 이유다. 올바른 빅데이터 해석을 위해선 샘플링 오류(sampling error)와 샘플링 편향(sampling bias)을 항상 염두에 둬야 한다. 한국정보화진흥원이 발간한 **<2016 디지털정보격차 실태조사>**에 따르면 여전히 장노년, 장애인, 농어촌 등 정보화 취약 계층이 존재한다. 카카오톡으로 퍼진 허위정보/가짜뉴스에 장노년층이 특히 많이 현혹되었던 사례가 있다. 같은 스마트폰을 사용하더라도 카카오톡만 이용할 줄 아는 세대와 다른 서비스를 이용하는 계층간의 정보 격차 역시도 존재하고 있다.

빅데이터 분석 결과를 전체 여론을 등치시키는 것은 항상 위험이 따른다. 그래서 늘 빅데이터 분석시에는 전통적인 통계관행을 잊지 말라는 말이 있다.
트위터에는 정치 관련 개인 견해들이 많이 올라온다. 트위터 빅데이터 분석 결과가 여론을 대표한다고 볼 수 있을까? 트위터를 사용하는 세대가 사실 장노년층에는 많지 않은 점

을 감안할 때 이는 샘플링 오류가 발생할 가능성 크다. 또한 자기 당 지지자들이 주로 사용하는 특정 커뮤니티의 견해가 대부분 주류 사회의 정치적 견해라고 할 수는 없을 것이다.



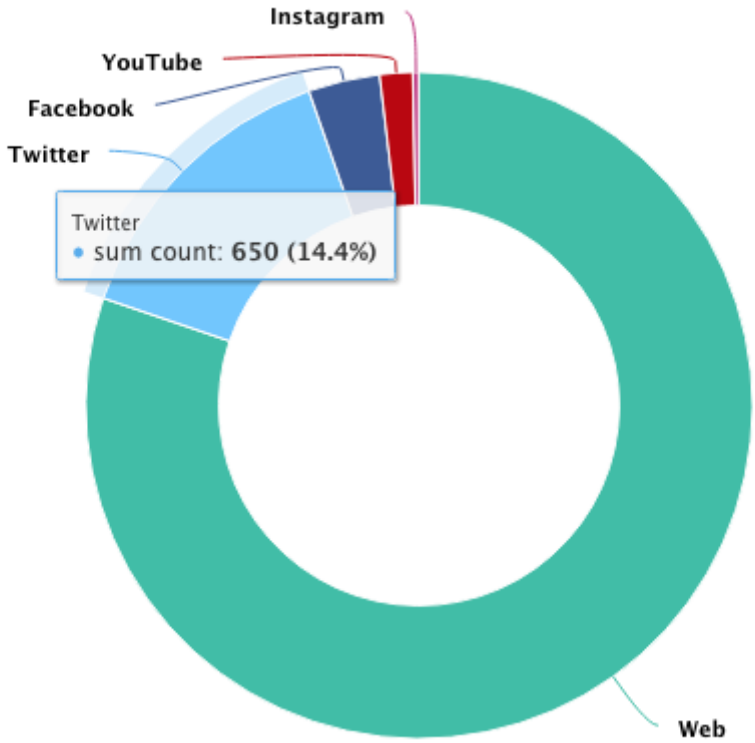
[2016 디지털정보격차 실태조사, **출처**]

데이터 샘플링과 편향성의 문제

<글로벌빅데이터연구소>에 따르면 "분석대상 채널은 누리꾼이 자신의 의견을 직간접으로 게재한 커뮤니티·블로그·카페·유튜브·트위터·인스타그램·페이스북·카카오스토리·지식인·기업/조직·정부/공공 등 모두 11개" 라고 밝혔다. 기간은 2020년 6월 3일~7일 오후 10시 30분까지 5일간 이재용 부회장에 대한 빅데이터를 대상으로 하였다고 적었다.

그런데 페이스북이나 인스타그램, 유튜브 등은 개인정보 보호문제로 인하여 공식적으로 빅데이터 업체들이 분석할 만한 양의 데이터를 API(Application Programming Interface)로 제공하지 않는다. 하여 편법인 크롤링이나 스크래핑을 통하여 개인 데이터를 소위 긁어야 한다. 특히 국내외 분석업체들은 페이스북 개인 데이터를 꼭 확보하고 싶어하나 실상은 뉴스 미디어나 브랜드 페이지 같은 외부 공개된 것들 이 외에는 접근이 쉽지 않다. 가짜 계정을 여러개 만들어 봇(Bot)을 통해 수집하는 편법적 방식이 아니라면 개인 포스팅 정보를 수집하는 것은 사실상 어렵거나 불가능하다. 그리고 시사 문제가 거의 올라오지 않고 이미지 콘텐츠 기반인 인스타그램은 분석대상 제외하는 것이 일반적이다.

아래 그림은 필자가 다양한 기법을 동원하여 "이재용" 키워드가 어떤 채널에서 언급됐는지 살펴본 결과다. 한달간 (5월 8일~6월 8일) 뉴스를 제외하고 이재용이 언급된 포스팅 전체의 채널별 분포다.



필자가 기사를 제외하고 한달간 '이재용'이란 키워드가 들어간 온라인의 포스팅 양을 채널별로 분석한 결과다. 글로벌빅데이터연구소의 자료에는 이런 기본적인 채널별 데이터 분포조차 없다.

위 그림을 보면, 웹이 압도적으로 많고, 그 뒤를 트위터, 페이스북, 유튜브가 따르고 있다. 웹과 트위터를 제외한 페이스북과 인스타그램에선 의견기후에 반영될만큼 충분한 양의 데이터를 모으는 것은 어렵다. 트위터는 상대적으로 데이터를 얻기 쉽다. 해당 데이터를

키워드 검색기반으로 서비스 해주는 회사들도 많다. 문제는 트위터 역시 노년층 사용률이 떨어져 국민 의견을 반영한다고 보기는 힘들다.

유튜브의 경우 영상 콘텐츠와 댓글들이 분석 대상이 될 것인데 영상의 경우 내용분석을 하기 전에는 알 수 없다. 영상에 달린 댓글의 경우 해당 영상이 보수성향인지 진보성향인지에 따라 같은 어휘라도 전혀 다른 의미로 쓰일 수 있어 분석에 주의를 기울여야 한다. 하지만 <글로벌빅데이터연구소> 연구에는 해당 내용에 대한 설명이 전혀 없다.

가장 높은 비중을 차지하는 웹은 온라인 커뮤니티와 포털의 카페„ 언론사 댓글 및 정부 사이트들이 대부분이다. 이 경우 보수와 진보 성향으로 나뉘어진 경우가 많다. 유유상종 효과가 강해 어떤 커뮤니티를 대상으로 했는지 명확히 밝혀줘야 한다. 만약에 일간베스트를 주로 분석한 뒤 이걸 국민 여론이라고 포장해서는 안된단 말이다. 그건 그냥 '일베 여론'인 것이다. 글로벌빅데이터연구소는 이런 샘플링의 투명성을 거의 지키지 않았다.

<글로벌빅데이터연구소>는 5일동안 '이재용' 이름이 거론된 총 게시물 수가 뉴스를 제외하고 4783건이라고 하였다. 그런데 필자가 확인한 총 게시물수는 트위터의 경우 한 달동안 654개(14.4%) 였다. 트위터 유저들은 주로 언론사들의 기사를 단순 리트윗하는 경우가 많아 뉴스와 함께 해당 뉴스를 리트윗한 글 역시 동일한 것으로 간주하고 추가적인 멘션이 없는 한 제외해야 한다.

필자의 데이터 수집량과 분포를 기준으로 보면 <글로벌빅데이터연구소> 데이터는 70% 이상이 출처가 불분명한 웹데이터라고 볼 수 밖에 없다. 달랑 4783건 분석을 가지고 국민 의견 운운하는 것은 어불성설이다. 참고로 필자가 모 기업의 특정 브랜드명 관련 소셜 빅데이터 분석을 할때 사용한 데이터량이 3000만건이었다. <글로벌빅데이터연구소> 분석이 얼마나 의미 없는지를 알 수 있다.

2. 내맘대로 긍정/부정으로 분류? 어휘가 쓰인 맥락을 봐야 한다

소셜미디어상의 이용자들이 올리는 글과 이미지등의 빅데이터 분석을 통해 인사이트를 찾는 과정을 소셜 모니터링(Social Monitoring) 혹은 소셜리스닝(Social Listening)이라고 한다. 사용자들의 올리는 글을 통해 얻을 수 있는 인사이트는 어떤 분석방법을 사용하느냐에 따라 달라진다. 해당 글에서 가장 많이 언급되는 단어를 분석하는 방법으로 워드클라우드나 TD-IDF가 있다. 해당 글의 감성(긍정, 부정, 중립, 모름)을 분석하는 방법은 감정 분위기 분석(Sentiment/Mood Analysis)이 있다. 이밖에 글의 주요 주제를 파악하기 위해선 토픽분석, 의미분석 및 구문 분석 등이 활용된다. 이를 위해선 수많은 자연어 처리 알고리즘들이 사용된다.

[소셜 빅데이터 분석단계]

문제는 <글로벌빅데이터연구소>가 사용한 방법론이 정확히 무엇인지 알 수 없다는 것이다. 글로벌빅데이터연구소는 '이재용'이 거론된 4783건 게시물을 분석했다고 말하며 "단어 기반 연관어 분석기법이 빅데이터상 국민들의 의견을 분석할 때 가장 유용한 방법 중 하나"라고 주장하고 있다. 그런데 이런 주장은 빅데이터 전문가인 필자도 처음 들어보는 말이다. 의견 기후를 분석하려면 앞서 언급했듯이 문장과 글의 의미분석/구문분석/토픽분석과 같은 자연어 처리가 더해져야 한다.

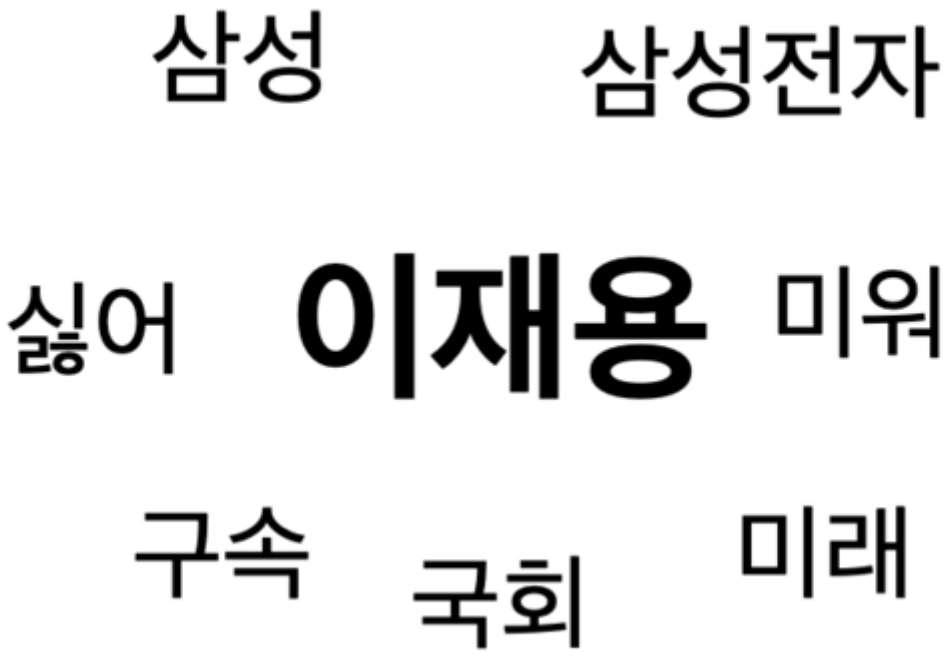
아래는 <글로벌빅데이터연구소>가 빅데이터(?) 4783건을 분석해 제시한 워드 클라우드다. 일반적인 다른 워드 클라우드와 비교해도 얼마나 허접한지 금방 알 수 있다.

[글로벌빅데이터연구소 분석자료 : **출처**]

워드 클라우드는 수집한 글들에서 가장 많이 언급된 단어를 중복회수 크기별로 나열한 것이다. 어떤 단어가 문장에나 글에서 많이 쓰였나 정도의 정보만 주는 것이지 단어간 관계를 명확히 설명할 수 없다. 단어간 관계는 워드클라우드로 하는 것이 아니라 워드임베딩(Word2Vec, BERT)같은 기법으로 진행한다. 워드 클라우드의 원리는 다음 예시를 보면

쉽게 이해할 수 있다. 필자가 아래의 다섯 문장 (삼성전자 이재용 미워, 이재용 싫어, 이재용 구속, 이재용을 국회로, 삼성의 미래는 이재용)을 웹에서 수집했다고 가정하고 이를 워드 클라우드로 그리면 아래와 같은 그림이 나온다.

"삼성전자 이재용 미워"
"이재용 싫어"
"이재용 구속"
"이재용을 국회로"
"삼성의 미래는 이재용"



[워드 클라우드는 딱 이정도의 인사이트만을 주는 것이다.]

이재용이 다섯번이나 언급됐으니 가장 크게 들어간 것이고 나머지 어휘들을 이재용 주변에 쭉 배치한 것이다. 그렇다면 문장 내에서 특정 단어가 계속 반복되면 무조건 중요하고 의미가 있다고 믿을 수 있을까? 그렇지 않다. TD-IDF(Term Frequency - Inverse Document Frequency) 라는 알고리즘이 있다. 쉽게 설명하면 특정 단어가 글에서 반복되면 그 단어는 중요하다고 생각하는 것이 일반적이다. 하지만 다른 글에서도 자주 등장하는 단어라면 중요도는 낮아질 수밖에 없다. 어쩌면 너무나도 당연한 논리이다. 단순 단어 빈도수만으로 중요도나 또 다른 의미를 찾을 수 없다는 것이다.

자연어처리, 의미 분석과 토픽 분석

텍스트 데이터로부터 의미 있는 정보를 찾아낼 수 있을 것이라는 기대는 통계적 의미론 가설 (**statistical semantics hypothesis**)에 근거한다. 텍스트 마이닝의 기반이 되는 이 가설은 "사람들이 쓰는 글이나 말에 등장하는 단어들의 통계적 규칙성으로부터 사람들이 말하고자 하는 의미와 정보를 찾아낼 수 있다"는 터니와 판텔의 연구 (**Turney and Pantel, 2010**)를 전제로 하고 있다. 이러한 연구들은 대부분 특정 키워드에 의존하는 방식을 택하고 있다. 다만 키워드 방식은 정확성에 한계가 있기 때문에 정확성을 검증할 만한 통계적인 지표를 제시해줘야 한다.

가장 어려운 것은 "선처"나 "불관용" 등 텍스트로부터 사람의 심리와 감정을 유추하는 것이다. 이를 위해선 엄청난 사전 데이터가 있어야 한다. 심리 분석이 얼마나 방대한 데이터를 필요로 하는지를 보여주는 사례가 있다. 2017년 서울대와 한국은행은 <텍스트 마이닝 기법을 이용한 경제심리 관련 문서 분류> **보고서**를 발표했다. 2007년 1월부터 2017년 6월까지 **10년간 한글로 작성된 경제 언론기사 전체**를 데이터셋으로 했음에도 아직 연구가 더 필요하다라고 결론을 내렸다. 딱 이 정도가 현재 텍스트 마이닝을 통한 심리분석 수준이다. 다시 <글로벌빅데이터연구소>의 분석자료를 보자.

이재용 부회장 검찰 구속영장 관련 연관어 추이

정보량 순위	연관 키워드	정보량	정보량 증가율
1	이재용	4634	185.2
2	삼성	3397	187.9
3	부회장	2835	192.6
4	삼성전자	2402	164.2
5	검찰	1985	460.7
6	기소	1013	NEW
7	삼성물산	964	305
8	의혹	954	265.5
9	경영권	942	275.3
10	사건	927	339.3
11	정부	897	120.9
12	제일모직	856	490.3
13	미국	814	94.3
14	심의위원회	783	NEW
15	경영	772	185.9
16	한국	767	65.3
17	서울	758	126.3
18	위기	752	124.5
19	국민	734	478.0
20	못한다	724	115.5
21	우려하다	697	155.3
22	문제	688	135.6
23	전략	673	158.8
24	전문가	656	375.4
25	세계	630	45.5
26	시장	621	37.7
27	생각	619	136.3
28	회사	607	78.5
29	미래	602	75.5
30	사태	588	117.8

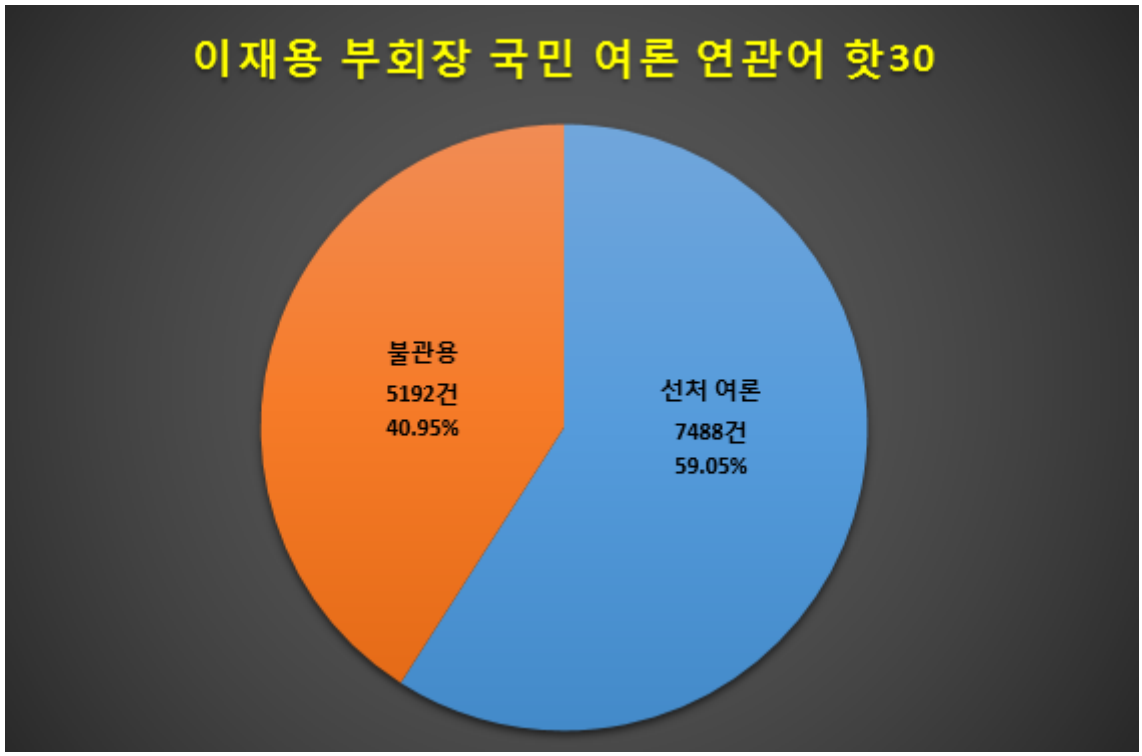
주) 1. 조사기간은 삼성 검찰수사심의위 신청일인 6월3일부터 7일까지 5일간임.
2. 정보량 비교 기간은 5월29~6월2일 5일간임.
3. 이들 연관어의 원문글을 일일이 확인한 결과 주황색은 불관용 의견, 녹색은 선처 의견 비중이 더 높았다.
4. 회색 의견은 가치판단이 배제된 중립어들임

글로벌빅데이터연구소가 제시한 이재용 검찰 구속영장 관련 연관어 추이

정보량은 해당 단어가 들어간 포스팅(글) 수다. 수집된 데이터 중에서 많이 나오는 단어를 빈도수 순으로 나열한 것이다. 정보량 증가율은 사실상 의미없는 자료다. 그리고는 연구소는 다음과 같이 설명한다.

선처의견 연관어를 구체적으로 살펴보면 '심의위원회' 783건, '경영' 772건, '한국' 767건, '국민' 734건, '우려하다' 697건 등이었다.
불관용 의견 연관어의 경우 '삼성물산' 964건, '의혹' 954건, '경영권' 942건, '제일모직' 856건, '위기' 752건 등이다."
예컨대 '국민' 키워드를 클릭, 원문들을 살펴보면 이재용 부회장에 대해 국민들의 의견을 구하는 글에서 결론이 선처의견이 많지만 불관용의견도 적지는 않다. 다만 과반수가 선처 의견인 것이다.

필자가 <글로벌빅데이터연구소>의 방법을 유추하면 다음과 같다. 포스팅글에서 예를 들면 '전문가'(위 도표에서 24번째 녹색 배경) 라는 단어가 들어간 글들을 수작업으로 직접 보니 과반수 이상이 '선처의견' 이었기 때문에 '전문가'가 들어간 문장은 대부분 '선처의견'으로 본 것이다. '경영권'(위 도표에서 9번째 녹색 배경) 이라는 단어가 들어간 글들을 수작업으로 직접 보니 과반수가 '불관용' 관련 글들이니까 '경영권'이 들어간 문장이나 글은 대부분 불관용으로 보겠다라는 말이다. 정말 한숨이 나올 지경이다.



[상당수의 언론사가 인용한 그래프 : **출처**]

자 이제 클라이맥스다. <글로벌빅데이터연구소>의 연구결론이자 대다수 언론사가 이재용 구속심사 직전에 열심히 퍼나른 도표다. 글로벌빅데이터연구소는 4783건의 데이터를 기준으로 분석했다고 했다. 샘플링 데이터가 4783건이라는 말이고 이는 중복자를 차치하고라도 4783명의 의견이라고 보는 것이다. 그럼 위의 도표는 합이 4783건(4783개의

의견표시글)이 되어야 하는데 합쳐보면 선처 7488건, 불관용 5192건으로 1만2680건이 나온다. 자세히 보면 선처 여론이라고 규정한 단어들이 포함된 문장수, 불관용 여론이라고 규정한 단어들이 포함된 문장수를 교집합이 아닌 합집합으로 중복해서 더해버렸다.

> "반도체 **전문가** 이재용은 **세계 시장** 진출을 위하여 깊은 **생각**과 고민을 통해 **한국**의 **미래를** 위하여 관련 **시장**에서 용기를 가지고 행동하는 **전문가**이다.

예를 들면 <글로벌빅데이터연구소> 분석에 따르면 이재용 선처를 주장한 위의 한 문장은 선처여론 7건이 되는 것이다.

3. 기업-분석업체-언론의 삼각 카르텔

이재용 부회장 구속영장 실질심사가 진행되는 과정에서 빅데이터로 포장된 이런 엉터리 분석이 나오고 이를 수백개 언론이 받아쓰는 현상을 어떻게 봐야할까. 유독 삼성 등 재벌 앞에서 한없이 비굴해지는 곳이 많은 이유는 구구절절 설명하지 않아도 알 것이다. 부끄러운 줄 알아야 한다. 기업-분석업체-언론의 신성동맹, 삼각 카르텔이라고 불러도 무방할 것이다.

<글로벌빅데이터연구소>의 홈페이지상 다른 보도자료들(빅데이터를 빙자한 기업홍보성 기사들)과 연구소장의 약력 및 연구위원이라는 비상근직 위원들만 봐도 사실 분석의 수준은 충분히 예측할 수 있었다. 그렇다고 빅데이터 분석이 대단한 학교의 IT전공자들만 할 수 있는 분야는 아니다. 다만 과학적인 방법을 통하여 합리적으로 정확한 사실만 전달하는 것은 생명으로 여겨야 하지 않을까?

이미 조선일보는 과거(**'핵포기 안해' 김정은 속마음 AI 분석? 조선일보의 '조작'**)에서 검증되지 않은 데이터분석회사를 끌어들여 김정은 속마음을 분석한다는 사기성 기사를 내보낸 사례가 있다. AI나 빅데이터 등 현란한 어휘로 포장만 하려 하지 말고 제대로 된 분석을 하길 바란다. 그리고 언론윤리에 대해서도 좀 더 고민을 하길 기대해본다. 팩트체크는 커녕 기초적인 내용도 확인하지 않고 그대로 받아쓰는 언론들은 게으른건지 삼성이라 일부러 그러는건지 궁금하다.

<글로벌빅데이터연구소>의 이번 분석은 거의 대국민 사기에 가깝다. 뉴스톱은 항상 반론에 열려 있다. 회사측의 반론을 기대한다. 참고로 필자가 트위터글만을 대상으로 분석했을 때에는 약 70%가 불관용이었다. 그렇다고 필자는 "국민 70%가 이재용 구속에 대하여

불관용 태도를 보이고 있다"고 말하지 않을 것이다. 데이터가 너무 적고 샘플링 편향성이 있을 수밖에 없기 때문이다.

지윤성 팩트체커